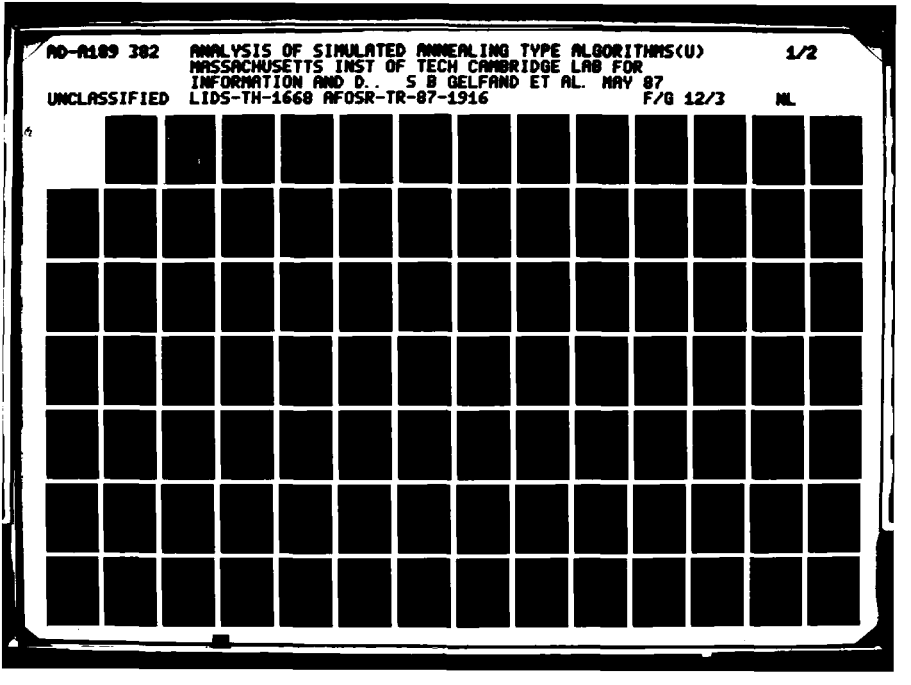
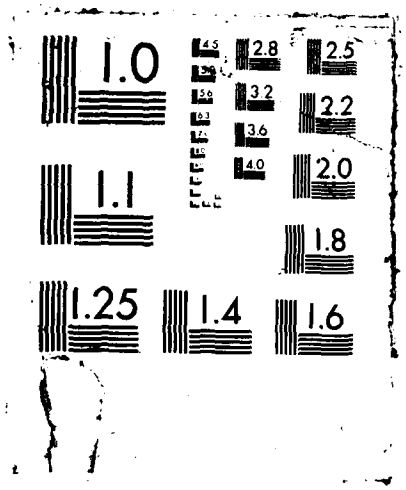
AD-A189 382 ANALYSIS OF SIMULATED ANNEALING TYPE ALGORITHMS(U) 1/2
MASSACHUSETTS INST OF TECH CAMBRIDGE LAB FOR
INFORMATION AND D.. S B GELFAND ET AL. MAY 87
UNCLASSIFIED LIDS-TH-1668 AFOSR-TR-87-1916 F/G 12/3 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
10.0

AD-A189 382

AFOSR-TR- 87-1916

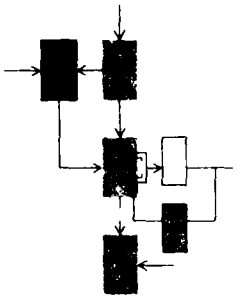
DTIC FILE COPY

MAY 1987

LIDS-TH-1668

Research Supported By:

AFOSR-85-0227
ARO DAAG29-84-K-0005
ARO DAAL03-86-G-0208



ANALYSIS OF SIMULATED ANNEALING TYPE ALGORITHMS

Saul B. Gelfand

DTIC
ELECTE
JAN 14 1988
S D
H

Laboratory for Information and Decision Systems

MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution [illegible]

87

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR. 87-1916
6a. NAME OF PERFORMING ORGANIZATION: Mass. Inst. of Tech; 6b. OFFICE SYMBOL (If applicable):	7a. NAME OF MONITORING ORGANIZATION: AFOSR/NM
6c. ADDRESS (City, State, and ZIP Code): Cambridge, Mass. 02139	7b. ADDRESS (City, State, and ZIP Code): AFOSR/NM, Bldg 410, Bolling AFB DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR; 8b. OFFICE SYMBOL (If applicable): NM	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: AFOSR-85-0227
8c. ADDRESS (City, State, and ZIP Code): AFOSR/NM, Bldg 410, Bolling AFB DC 20332-6448	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
61102	2304	A1	

11. TITLE (Include Security Classification)

Analysis of Simulated Annealing Type Algorithms

12. PERSONAL AUTHOR(S)

Saul B. Gelfand, Sanjoy K. Mitter

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Journal	FROM _____ TO _____	May 1987	103

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
Maj. James Crowley	(202) 767-5025	NM

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

May 1987 LIDS-TH-1668

ANALYSIS OF SIMULATED ANNEALING TYPE ALGORITHMS

by

Saul B. Gelfand

This report is based on the unaltered thesis of Saul B. Gelfand submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. This research was carried out at the MIT Laboratory for Information and Decision Systems with support provided in part by the following grants: AFOSR-85-0227, ARO DAAG29-84-K-0005, and ARO DAAL03-86-G-0208.

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist Avail and/or Special
A-1

Massachusetts Institute of Technology
Laboratory for Information and Decision Systems
Cambridge, MA 02139

ANALYSIS OF SIMULATED ANNEALING TYPE ALGORITHMS

by

Saul B. Gelfand

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS OF THE
DEGREE OF

DOCTOR OF PHILOSOPHY
IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
February 1987
© Massachusetts Institute of Technology 1987

Signature of Author _____
Department of Electrical Engineering and Computer Science
February 6, 1987

Certified by _____
Sanjoy K. Mitter
Thesis Supervisor

Accepted by _____
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

ANALYSIS OF SIMULATED ANNEALING TYPE ALGORITHMS

by

SAUL B. GELFAND

Submitted to the Department of Electrical Engineering and Computer Science
on February 6, 1987 in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy in
Electrical Engineering and Computer Science

ABSTRACT

The annealing algorithm is a popular Monte-Carlo algorithm for combinatorial optimization. The annealing algorithm consists of simulating a nonstationary finite state Markov chain whose state space is the domain of the cost function, called energy, to be minimized. The degree of randomization in the annealing algorithm is controlled by a parameter, called temperature, which is slowly decreased to zero. The convergence in probability and the rate of convergence of the annealing chain for the special case of an energy function with two local minima is analyzed. The sample path properties of annealing chains (with arbitrary energy functions) are examined. A modification of the annealing algorithm which makes noisy measurements of the energy function is given. The annealing algorithm is extended for optimization on general spaces.

The Langevin algorithm is a popular Monte-Carlo algorithm for multivariate optimization. The Langevin algorithm consists of simulating a nonstationary diffusion process. The relationship between the annealing and Langevin algorithms is studied. It is shown that an annealing chain driven by white Gaussian noise and interpolated into a piecewise constant process converges weakly to a time-scaled Langevin diffusion. Motivated by this result, a hybrid annealing/Langevin algorithm is proposed.

Thesis Supervisor: Dr. Sanjoy K. Mitter

Title: Professor of Electrical Engineering

this is dedicated
to my mom and dad

ACKNOWLEDGMENTS

I would like to thank Professor Sanjoy Mitter for suggesting the topic of this research, and for his technical assistance, financial support, and personal encouragement throughout the course of my graduate program. I would particularly like to thank him for understanding my need to take a systematic approach to learning, even when this might not have been the quickest way to get results, for it is my hope and belief that this route will pay rich dividends in the future.

I am also grateful to Professors Bruce Hajek and Lennart Ljung for some useful discussions on simulated annealing and related topics, as well as the members of my thesis committee Professors Dimitri Bertsekas, Thomas Magnanti, and Vivek Borkar.

Finally, I would like to thank my fellow graduate students and good friends Jerome Abernathy, Yehuda Avniel, and David Flamm for many enjoyable technical and non-technical discussions alike. But most of all I must thank my parents for their selfless unswerving support.

This research was supported in part by the following grants: AFOSR-85-0227, DAAG19-84-K-0005, and DAAL03-86-G-0208.

TABLE OF CONTENTS

Page

FREQUENTLY USED NOTATION

$\mathbb{N}$, the natural numbers

$\mathbb{R}^r$, r-dimensional Euclidean space

$\mathbb{B}^r$, Borel subsets of $\mathbb{R}^r$

$C^r[0,T]$, $\mathbb{R}^r$-valued continuous functions on [0,T]

$D^r[0,T]$, $\mathbb{R}^r$-valued càdlàg functions on [0,T]

$N(m,\Lambda)$ $(\cdot)$, Normal measure with mean m and covariance Λ

$\chi_A(\cdot)$, indicator function of the set A

B(a,R), ball of radius R centered at a

$|x|$, Euclidean norm of x

(x,y), Euclidean inner product of x,y

$x \otimes y$, Euclidean outer product of x,y

a.e., almost everywhere

w.p.1, with probability one

i.o.,, infinately often

a.a., almost always

If $\{a_k\}$ and $\{b_k\}$ are sequences of real numbers with $a_k \neq 0$ for k large enough then

$$b_k = O(a_k) \quad \text{if} \quad \limsup_{k\to\infty} \left|\frac{b_k}{a_k}\right| < \infty$$

$$b_k = o(a_k) \quad \text{if} \quad \lim_{k\to\infty} \frac{b_k}{a_k} = 0$$

$$b_k \sim a_k \quad \text{if} \quad \lim_{k\to\infty} \frac{b_k}{a_k} = 1$$

CHAPTER I
INTRODUCTION

Algorithms for finding a global extremum of a real-valued function may be classified into two groups: deterministic and random. The distinction here is of course that the random or Monte-Carlo algorithms make use of pseudo random variates whereas the deterministic algorithms do not. The earliest global optimization algorithms were of the deterministic type and were associated with evaluating the cost function at points on a grid. One drawback of these methods is that they typically require certain prior information about the cost function such as a Lipshitz constant. Most global optimization algorithms are of the random type and are related to the so-called multistart algorithm. In this approach, a local optimization algorithm is run from different starting points which are selected at random, usually from a uniform distribution on the domain of the cost function. See [5], [29] for a discussion of global optimization algorithms.

Recently, motivated by hard combinatorial optimization problems such as arise in computer design and operations research, Kirkpatrick et. al. [19] and independently Cerny [3] have proposed a different kind of random algorithm called *simulated annealing*. The annealing algorithm is based on an analogy between large scale optimization problems and statistical mechanics. For our purposes this analogy consists simply of viewing the cost function as an energy function defined on a finite state space of an imaginary physical system. The annealing algorithm is then seen as a variation on a Monte-Carlo algorithm developed by Metropolis et. al. [25] for making statistical mechanics calculations, which we now describe. It is well-known that the states of a physical system in thermal equilibrium obey a Gibbs distribution $\propto \exp[-U(\cdot)/T]$, where $U(\cdot)$ is an energy function and T is the temperature. The Metropolis algorithm was developed for obtaining samples from such a Gibbs distribution and for computing estimates of functionals averaged over the Gibbs distribution. The Metropolis algorithm proceeds as follows:

> Given a state i of the system, select a candidate state j in a random manner corresponding to a small perturbation of the system, and

compute the change in energy $\Delta U = U(j) - U(i)$. If $\Delta U \leq 0$ accept state j as the new state for the next iteration of the algorithm. If $\Delta U > 0$ accept state j with probability $\exp[-\Delta U/T]$; otherwise the algorithm starts at state i for the next iteration.

The annealing algorithm consists of identifying the cost function to be minimized with the energy function $U(\cdot)$ and taking the temperature T as a function of time and slowly lowering it to zero. Suppose that the distribution of a candidate state is independent of past states given the current state. Then it is clear that the Metropolis algorithm simulates the sample paths of a Markov chain, and it can be shown that if the candidate states are selected in a suitable manner then this chain infact has a Gibbs distribution $\propto \exp[-U(i)/T]$ as its (unique) equilibrium distribution (see Chapter 2 for details). Furthermore as the temperature T is decreased to zero the Gibbs distribution concentrates more and more on the lower energy states. The motivation behind the annealing algorithm is that if $T \rightarrow 0$ slowly enough such that the system is never far away from equilibrium, then presumably there is convergence (in some probabilistic sense) to the global minima of $U(\cdot)$.

The annealing algorithm stands in contrast to heuristic methods for combinatorial optimization which are based on iterative improvement, allowing only decreases in the cost function at each iteration. Iterative improvement algorithms in statistical mechanics terms correspond to rapidly quenching a system from a high to a very low temperature. Such quenching can result in the system getting trapped in a so-called metastable state, and analogously the iterative improvement algorithm getting trapped in a strictly local minimum of the cost function. On the other hand, the annealing algorithm corresponds to slowly cooling a system. Such cooling should result in the system spending most of its time among low energy states and analogously the annealing algorithm finding a global or nearly global minimum of the cost function.

The annealing algorithm as described above is suitable for combinatorial optimization. Motivated by optimization problems with continuous variables which arise in image processing problems, Geman and independently Grenander [13] have proposed a diffusion-type algorithm called the *Langevin algorithm* (as coined by Gidas [11]). Consider the diffusion solution of the Langevin equation

$$dx(t) = -\nabla U(x(t))dt + \sqrt{2T}\, dw(t)$$

where $U(\cdot)$ is now a smooth function on r-dimensional Euclidean space (again called energy), T is a positive constant (again called temperature), and $w(\cdot)$ is

a standard r-dimensional Wiener process. The Langevin equation describes the motion of a particle in a viscous fluid. The Langevin algorithm consists of identifying the cost function to be minimized with the energy function $U(\cdot)$ and taking the temperature T as a function of time and slowly lowering it to zero. Now it is well known that under suitable conditions on $U(\cdot)$ the diffusion solution of the Langevin equation has a Gibbs density $\propto \exp[-U(\cdot)/T]$ as its (unique) equilibrium density, and as the temperature T is decreased to zero this density becomes more and more concentrated on the lower energy states. Like the annealing algorithm, the motivation behind the Langevin algorithm is that if $T \rightarrow 0$ slowly enough such that the system is never far away from equilibrium, then presumably there is convergence (in some probabilistic sense) to the global minima of $U(\cdot)$.

The annealing algorithm has been applied with varying success to a wide range of problems including circuit placement and wire routing for VLSI chip design [19], image reconstruction [8], and assorted hard combinatorial problems which arise in operations research [3], [12], [18], [19]. There has also been intense theoretical interest in both the annealing algorithm [8], [10], [11], [14], [15], [26], [31] and the Langevin algorithm [4], [9], [11], [15], [21].

The goal of this thesis may simply be stated as the analysis of the asymptotic (large time) behavior of simulated annealing type algorithms, by which we mean not only the annealing algorithm but also the Langevin and related algorithms. We are particularly interested in the relationship between the annealing and Langevin algorithms. Here is a Chapter-by-Chapter outline of the thesis.

In Chapter 2 we discuss the finite state annealing algorithm as proposed by Kirkpatrick and independently by Cerny. In 2.1 we give a precise description of the annealing chain (the Markov chain whose sample paths are simulated in the annealing algorithm). We then briefly discuss two numerical studies of the annealing algorithm by Johnson et. al. [18] and Golden and Skiscim [12], and next describe some of the large body of theoretical work on the subject with particular emphasis on the work of Mitra et. al. [26] and Hajek [14]. In 2.2 we study the asymptotic behavior of a class of nonstationary finite state Markov chains in preparation for the analysis of the annealing algorithm itself. In 2.3 we use the results of 2.2 to analyze the annealing algorithm. We first examine in depth the convergence in probability and the rate of convergence of the annealing chain to the globally minimum energy state for an energy function with two local minima (one strictly local and one global). Although cost functions encountered in large scale combinational problems may have large numbers of local minima, the

results we present are new and offer some interesting insights. We next perform a sample path analysis of the annealing chain and obtain conditions under which the annealing chain visits the set of global minima of the energy function with probability one, visits the set of global minima with probability strictly less than one, or converges to the set of global minima with probability one. These results are different than most of the analytical results on the annealing algorithm, which give conditions under which the annealing chain converges to the set of global minima in probability. In 2.4 we describe and analyze a modification of the annealing algorithm which uses noisy measurements of the energy function.

In Chapter 3 we extend the annealing algorithm for optimization on general spaces. In 3.1 we give a precise description of a general state annealing chain. In 3.2 we discuss the ergodicity of the general state annealing chain at a fixed temperature, i.e., we discuss a general state version of the Metropolis algorithm. Here we settle some technical issues which do not arise in the finite state Metropolis algorithm. In 3.3 we study the asymptotic behavior of a class of nonstationary general state Markov chains in preparation for the analysis of the general state annealing algorithm itself. In 3.4 we use the results of 3.3 to extend the result of 2.3 on the finite state annealing chain visiting the set of global minima of the energy function with probability one to the general state case, essentially under the conditions that the state space be a compact metric space and the energy function be continuous. It is not known whether convergence to the set of global minima in probability can be obtained under such weak conditions.

In Chapter 4 we discuss the Langevin algorithm as proposed by Geman and independently by Grenander. In 4.1 we give a precise description of the Langevin algorithm and summarize the convergence results of Geman and Hwang [9], Gidas [11], and Kushner [21]. In 4.2, 4.3 we present what we believe to be the most interesting results of the thesis. In 4.2 we show that an annealing chain of the type considered in Chapter 3 with r-dimensional Euclidean state space and driven by white Gaussian noise converges in a certain sense to a Langevin diffusion. In 4.3 we propose a hybrid annealing/Langevin algorithm based on the results of 4.2. We argue that the hybrid algorithm enjoys the advantages of both the annealing and Langevin algorithms. Unfortunately, we have not yet succeeded in establishing the convergence of the hybrid algorithm and this is left as a future task.

In Chapter 5 we collect the results of the thesis and make some concluding remarks.

CHAPTER II
FINITE STATE ANNEALING TYPE ALGORITHMS

2.1 Introduction to the Annealing Algorithm

In Chapter 1 we briefly described the annealing algorithm and discussed the heuristic motivation based on the connection that Kirkpatrick [19] has suggested between statistical mechanics and large-scale optimization problems. Mathematically, the annealing algorithm consists of simulating a nonstationary finite-state Markov chain whose state space is the domain of the cost function (called energy) to be minimized. In this Section we shall discuss in detail the annealing algorithm and describe some of the considerable literature which has been devoted to its analysis.

We first give some standard finite state space Markov chain notation (c.f. [6], [7]). Let Σ be a finite set. $P = [p_{ij}]_{i,j\in\Sigma}$ is a stochastic matrix on Σ if $p_{ij} \geq 0$ for all $i,j\in\Sigma$ and

$$\sum_{j\in\Sigma} p_{ij} = 1 \qquad \forall\, i\in\Sigma\,.$$

$\{P^{(k,k+1)}\} = \{[p_{ij}^{(k,k+1)}]\}$ are the 1-step transition matrices for a Markov chain $\{\xi_k\}$ with state space Σ if for every $k\in\mathbb{N}$ $P^{(k,k+1)}$ is a stochastic matrix on Σ and

$$P\{\xi_{k+1} = j \,|\, \xi_k = i\} = p_{ij}^{(k,k+1)} \quad (\text{if } P\{\xi_k = i\} > 0) \tag{2.1}$$

for all $i,j\in\Sigma$. Conversely, given a sequence $\{P^{(k,k+1)}\} = \{[p_{ij}^{(k,k+d)}]\}$ of stochastic matrices on Σ we can construct on a suitable probability space (Λ,F,P) a Markov chain $\{\xi_k\}$ with state space Σ which satisfies (2.1). For each $d\in\mathbb{N}$ let

$$P^{(k,k+d)} = P^{(k,k+1)} \cdot \ldots \cdot P^{(k+d-1,k+d)}\,.$$

$P^{(k,k+d)} = [p_{ij}^{(k,k+d)}]$ is a stochastic matrix on Σ and

$$P\{\xi_{k+d} = j \,|\, \xi_k = i\} = p_{ij}^{(k,k+d)} \quad (\text{if } P\{\xi_k = i\} > 0)$$

for all $i,j\in\Sigma$. It will be convenient to have a fixed version of the conditional probability of ξ_{k+d} given ξ_k which we define by

$$P\{\xi_{k+d} \in A \mid \xi_k = i\} = \sum_{j \in A} p_{ij}^{(k,k+d)}$$

for all $i \in \Sigma$ and $A \subset \Sigma$.

We now define the annealing algorithm. Let $U(\cdot)$ be a nonnegative function on Σ, called the *energy function*. The goal is to find a point in Σ which minimizes or nearly minimizes $U(\cdot)$. Let $\{T_k\}$ be a sequence of positive numbers, called the *temperature schedule*. Let $Q = [q_{ij}]$ be a stochastic matrix on Σ. Now let $\{\xi_k\}$ be the Markov chain with state space Σ and 1-step transition matrices $\{P^{(k,k+1)}\} = \{[p_{ij}^{(k,k+1)}]\}$ given by

$$p_{ij}^{(k,k+1)} = \begin{cases} q_{ij} \exp\left[-\dfrac{U(j) - U(i)}{T_k}\right] & \text{if } U(j) > U(i) \\ q_{ij} & \text{if } U(j) \leq U(i),\ j \neq i \\ 1 - \sum\limits_{\ell \neq i} p_{i\ell}^{(k,k+1)} & \text{if } j = i \end{cases} \tag{2.2}$$

for all $i,j \in \Sigma$. $\{\xi_k\}$ shall be called the *annealing chain*. For each $d \in \mathbb{N}$ let $Q^d = [q_{ij}^{(d)}]$. Recall that Q is *irreducible* if for every $i,j \in \Sigma$ there exists a $d \in \mathbb{N}$ such that $q_{ij}^{(d)} > 0$. Also, Q is *symmetric* if $q_{ij} = q_{ji}$ for all $i,j \in \Sigma$. In the special case where Q is irreducible and symmetric and $T_k = T$, a positive constant, $\{\xi_k\}$ is the stationary Markov chain introduced by Metropolis et. al. [25] for computing statistics of a physical system in thermal equilibrium at temperature T. It was Kirkpatrick et. al. [19] and Cerny [3] who suggested that the Metropolis scheme could be used for minimizing $U(\cdot)$ by letting $T = T_k \to 0$. We shall call the algorithm which simulates the sample paths of $\{\xi_k\}$ with $T_k \to 0$ the *annealing algorithm*.

The heuristic motivation behind the annealing algorithm was discussed (briefly) in Chapter 1. Here we give the motivation in more mathematical terms. Suppose that Q is irreducible and symmetric, and let $\{\xi_k^T\}$ be the stationary chain with 1-step (stationary) transition matrix $P^T = [p_{ij}^T]$ given by the r.h.s of (2.2) with $T_k = T$, a positive constant. Then it can be shown that P^T has an invariant Gibbs vector $\Pi^T = [\pi_i^T]$ (a row vector), i.e.,

$$\Pi^T = \Pi^T P^T$$

where

$$\pi_i^T = \frac{\exp\,[-\,U(i)/T]}{\sum\limits_{j\in\Sigma} \exp\,[-\,U(j)/T]} \qquad \forall\, i\in\Sigma\,.$$

This follows from the detailed reversibility

$$\pi_i^T p_{ij}^T = \pi_j^T p_{ji}^T \qquad \forall\, i,j\in\Sigma\ .$$

Furthermore, Q irreducible and symmetric implies that $\{\xi_k^T\}$ is an irreducible† (and aperiodic) chain and by the Markov Convergence Theorem [6, p. 177]

$$\lim_{k\to\infty} P\{\xi_k^T = i\} = \pi_i^T \qquad \forall\, i\in\Sigma\,. \tag{2.3}$$

Let S be the set of global minima of $U(\cdot)$, i.e.

$$S = \{i\in\Sigma:\ U(i) \le U(j)\ \ \forall\, j\in\Sigma\}\,.$$

Now

$$\lim_{T\to 0} \pi_i^T = \pi_i^* \qquad \forall\, i\in\Sigma \tag{2.4}$$

where $\Pi^* = [\pi_i^*]$ is a probability vector with support in S. In view of (2.3) and (2.4) the idea behind the annealing algorithm is that by choosing $T = T_k \to 0$ slowly enough hopefully

$$P\{\xi_k = i\} \approx \pi_i^{T_k} \qquad (k \text{ large}) \tag{2.5}$$

and then perhaps

$$\lim_{k\to\infty} P\{\xi_k = i\} = \pi_i^* \qquad \forall\, i\in\Sigma \tag{2.6}$$

and consequently ξ_k converges in probability to S.

In Chapter 1 we roughly described the procedure by which the sample paths of the annealing chain are simulated. It is seen that the Q matrix governs the small perturbations in the system configurations which are then accepted or rejected probabilistically depending on the corresponding energy changes and the temperature. More precisely, the annealing chain may be simulated as follows. Suppose $\xi_k = i$. Then generate a Σ-valued random variable η with $P\{\eta = j\} = q_{ij}$. Suppose $\eta = j$. Then set

†A stationary chain is irreducible if its 1-step (stationary) transition matrix is irreducible.

$$\xi_{k+1} = \begin{cases} j & \text{if } U(j) \leq U(i) \\ j & \text{if } U(j) > U(i) \text{ with probability } \exp\left[-\dfrac{U(j) - U(i)}{T_k}\right] \\ i & \text{else} \end{cases}$$

There are two in depth numerical studies of simulated annealing of which we are aware. Johnson et. al. [18] applied the annealing algorithm to four well-studied problems in combinational optimization: graph partitioning, number partitioning, graph coloring, and the travelling salesman problem. They compare the annealing algorithm with the best of the traditional algorithms for each problem. They found that although annealing is able to produce quite good solutions on three of the four problems, only on one of the four (graph partitioning) does it outperform the best of its rivals. Golden and Skiscim [12] have tested the annealing algorithm on routing and location problems, specifically the travelling salesman problem and the p-median problem. They conclude that there are more efficient and effective heuristics for these problems.

We shall now outline the convergence results on the annealing algorithm which are known to us. We refer the reader to the specific papers for full details.

Geman and Geman [8] were the first to obtain a convergence result for the annealing algorithm. The consider a version of the annealing algorithm which they call the *Gibbs sampler*. They show that for temperature schedules of the form

$$T_k = \frac{c}{\log k} \qquad (k \text{ large})$$

that if c is sufficiently large then (2.6) is obtained.

Gidas [10] also considers the convergence of the annealing algorithm and similar algorithms based on Markov chain sampling methods related to the Metropolis method.

We next discuss the work of Mitra et. al. [26]. The idea behind their work is similar to that of Geman and Geman and also Gidas in that they show that for temperature schedules which vary slowly enough the annealing chain reaches "quasiequilibrium", i.e., something like (2.5) holds. In order to state Mitra et. al.'s result we will need the following notation. Let

$$N(i) = \{j \in \Sigma : \ q_{ij} > 0\} \qquad \forall\, i \in \Sigma\ .$$

Let S_M be the set of states that are local maxima of $U(\cdot)$, i.e.,

$$S_M = \{i \in \Sigma : \ U(i) \geq U(j) \quad \forall\, j \in N(i)\}\ .$$

Let

$$r = \min_{i \in \Sigma \setminus S_M} \max_{j \in \Sigma} d(i,j)$$

where $d(i,j)$ is the minimum number of steps to get from state i to state j. Finally, let

$$L = \max_{i \in \Sigma} \max_{j \in N(i)} |U(j) - U(i)|\ .$$

Here is Mitra et. al.'s result:

Theorem 2.1 (Mitra et. al. [26]) Assume Q is irreducible and symmetric†. Let $T_k \downarrow 0$ and

$$\sum_{k=1}^{\infty} \exp\left(-\frac{r\,L}{T_{kr-1}}\right) = \infty\ . \tag{2.7}$$

Then

$$\lim_{k \to \infty} P\{\xi_k = i\} = \pi_i^* \qquad \forall\, i \in \Sigma\ . \tag{2.8}$$

Remarks

(1) If $T_k = c/\log k$ then (2.7) holds iff $c \geq r\,L$.

(2) An estimate of the rate of convergence in (2.8) is obtained for annealing schedules of the form $T_k = c/\log k$ for $c \geq r\,L$. Let

$$w = \min_{i \in \Sigma} \min_{j \in N(i)} q_{ij}\ ,$$

$$\gamma = \min_{i \in \Sigma \setminus S} U(i) - \min_{j \in S} U(j)\ .$$

It is shown that

$$P\{\xi_k = i\} = \pi_i^* + O\left(\frac{1}{k^{\min\{\alpha,\gamma\}}}\right) \qquad \text{as} \quad k \to \infty \tag{2.9}$$

where

†for just $q_{ij} > 0$ iff $q_{ji} > 0$ for all $i,j \in \Sigma$

$$\alpha = \frac{w^r}{r^r L/c} \quad , \quad \beta = \frac{\gamma}{c} \quad .$$

Since α and β are increasing and decreasing respectively with increasing c, it is suggested that $c \geq r\,L$ be chosen to maximize $\min\{\alpha,\beta\}$.

We next discuss the work of Hajek [14]. The idea behind his work is that for temperature schedules which vary slowly enough, the annealing chain escapes from local minima of $U(\cdot)$ at essentially the same rate as for a constant temperature. In order to state Hajek's result we will need the following notation. We shall say that given states i and j, i can *reach* j if there exists a sequence of states $i = i_0,\ldots,i_p = j$ such that $q_{i_n i_{n+1}} \geq 0$ for all $n = 0,\ldots,p-1$; if $U(i_n) \leq E$ (a nonnegative number) for all $n = 0,\ldots,p$ then we shall say that i can *reach* j *at height* E. We shall say that the annealing chain is *strongly irreducible* if i can reach j for all $i,j \in \Sigma$. Clearly, strong irreducibility is equivalent to Q irreducible, but we introduce strong irreducibility to conform with Hajek's notation. We shall also say that the annealing chain is *weakly reversible* if for every $E > 0$, i can reach j at energy E iff j can reach i at energy E, for all $i,j \in \Sigma$. Let S_m be the states that are local minima of $U(\cdot)$, i.e.,

$$S_m = \{i \in \Sigma : \; U(i) \leq U(j) \qquad \forall\, j \in N(i)\} \quad .$$

For each $i \in S_m \backslash S$ let $\Delta(i)$ be the smallest number E such that i can reach some $j \in \Sigma$ with $U(j) < U(i)$ at height $U(i) + E$. $\Delta(i)$ is the "depth" of the local (but not global) minimum i. Let

$$\Delta^* = \max_{i \in S_m \backslash S} \Delta(i) \, . \tag{2.10}$$

Here is Hajek's result:

Theorem 2.2 (Hajek [14]) Assume that the annealing chain is strongly irreducible and weakly reversible. Let $T_k \downarrow 0$. Then

$$\lim_{k \to \infty} P\{\xi_k \in S\} = 1 \tag{2.11}$$

iff

$$\sum_{k=1}^{\infty} \exp\left[-\frac{\Delta^*}{T_k}\right] = \infty \, . \tag{2.12}$$

Remark If $T_k = c/\log k$ then (2.12) and hence (2.11) holds iff $c \geq \Delta^*$. For this reason Δ^* has been called the *optimal constant* and $T_k = \Delta^*/\log k$ the *optimal schedule*.

We should also mention that Tsitsiklis [30] has proved of generalization of Theorem 2.2 which does not assume weak reversibility, using (and extending) the theory of singularly perturbed Markov chains.

In view of Theorem 2.2 and the refinement in [30], the analysis of the convergence in probability of the annealing algorithm is essentially complete, with the exception that it does not appear that anyone has determined the rate of convergence for optimal or nearly optimal temperature schedules. Recall that Mitra et. al. have shown that (2.9) holds if

$$T_k = \frac{c}{\log k} \quad , \qquad c \geq r\,L \quad ,$$

but $r\,L$ is in general much larger than Δ^*. In 2.2, 2.3 we shall analyze the rate of convergence in probability of the annealing algorithm for a special case with two local minima. We will obtain results on the convergence rate for nonparametric temperature schedules (schedules *not* of the form $T_k = c/\log k$) and also for temperature schedules $T_k = c/\log k$ for $c \geq \Delta^*$. We remark that in the latter case with $c = \Delta^*$ there is apparently some interesting and unexpected behavior. Our results are different although consistent with (2.9).

Also in 2.2, 2.3 we shall explore the sample path behavior (as opposed to the ensemble behavior) of the annealing algorithm. We shall give a number of results, the most important of which is conditions such that the annealing chain visit the set S (infinitely often) with probability one. Suppose we let

$$\varsigma_1 = \xi_1$$

$$\varsigma_{k+1} = \begin{cases} \xi_{k+1} & \text{if } U(\xi_{k+1}) < U(\varsigma_k) \\ \varsigma_k & \text{else .} \end{cases}$$

Note that if $\{\xi_k\}$ visits S with probability one then $\{\varsigma_k\}$ traps in S with probability one, and furthermore no additional evaluations of $U(\cdot)$ are required to compute $\{\varsigma_k\}$ over what are required to simulate $\{\xi_k\}$. Hence by just doubling the memory requirements and keeping track of $\{\varsigma_k\}$, it seems sufficient to show that $\{\xi_k\}$ visit S with probability one rather than converge to S in probability. Now it might be imagined that the conditions on the temperature schedule under which $\{\xi_k\}$ visits S with probability one are

weaker than those under which $\{\xi_k\}$ converges to S in probability. However, the proof of Theorem 2.2 shows that (assuming strong irreducibility and weak reversibility) $\{\xi_k\}$ visits S with probability one iff (2.12) holds. From this point of view our result does not offer anything new; infact the temperature schedules we consider are not even optimal. However, we believe our result is important in the following sense. In Chapter 3 we extend the annealing algorithm to general state spaces. It turns out that our result on the finite state annealing chain visiting S infinitely often with probability one can also be extended, essentially under the condition that the state space be a compact metric space and the energy function be continuous. It is not clear whether convergence to S in probability can be shown in such a general setting; the methods used to analyze the finite state case (quasiequilibrium distributions, large deviations and perturbation theory) do not seem directly applicable.

Finally, in 2.4 we give a modification of the annealing algorithm which allows for noisy measurements of the energy function and examine its convergence.

2.2 Asymptotic Analysis of a Class of Nonstationary Markov Chains

In this Section we analyze the asymptotic properties of a certain class of nonstationary (finite state) Markov chains. These chains will have the property that their 1-step transition probabilities will satisfy bounds similar to those satisfied by the d-step transition probabilities of the annealing chain. The results of this Section will be used in 2.3 to deduce corresponding asymptotic properties of the annealing chain.

We shall consider the following class of Markov chains. Let Σ be a finite set. Let $\alpha_{ij}, \beta_{ij} \in [0,\infty]$ for $i,j \in \Sigma$, and $\{\theta_k\}$ a sequence of real numbers with $0 < \theta_k \leq 1$. Let $\{\xi_k\}$ be a Markov chain with state space Σ and 1-step transition matrices $\{P^{(k,k+1)}\} = \{[p_{ij}^{(k,k+1)}]\}$ with the following property: there exists positive numbers A, B such that

$$p_{ij}^{(k,k+1)} \geq A\ \theta_k^{\alpha_{ij}} \tag{2.13}$$

$$p_{ij}^{(k,k+1)} \leq B\ \theta_k^{\beta_{ij}} \tag{2.14}$$

for all $i,j \in \Sigma$. Actually, we shall assume that (2.13) and/or (2.14) hold depending on the result we wish to prove.

2.2.1 Convergence in Probability and Rate of Convergence for a Three State System

We now establish the convergence in probability and rate of convergence of a Markov chain $\{\xi_k\}$ with state space Σ which satisfies (2.13) and (2.14) for a special case with $|\Sigma| = 3$. In 2.3.2 we shall apply this result to the annealing chain with an energy function which has two local minima. It will be useful here to consider the more detailed bounds

$$A_{ij}\theta_k^{\alpha_{ij}} \le p_{ij}^{(k,k+1)} \le B_{ij}\theta_k^{\beta_{ij}} \qquad \forall\, i,j\in\Sigma \ , \tag{2.15}$$

where A_{ij}, B_{ij} are positive constants. Here is our theorem.

Theorem 2.3 Let $\Sigma = \{1,2,3\}$ and assume that (2.15) holds. Let

$$a = \max\{\alpha_{21}, \alpha_{31}\} < \infty \ ,$$
$$b = \min\{\beta_{12}, \beta_{13}\} > a \ ,$$
$$\gamma = b - a \ ,$$
$$\delta = \begin{cases} \min\{A_{21}, A_{31}\} & \text{if } \alpha_{21} = \alpha_{31} \\ A_{21} & \text{if } \alpha_{21} > \alpha_{31} \\ A_{31} & \text{if } \alpha_{21} < \alpha_{31} \ . \end{cases}$$

(a) Suppose that $\theta_k \downarrow 0$ and

$$\sum_{k=1}^{\infty} \theta_k^a = \infty \ . \tag{2.16}$$

Then

$$\lim_{k\to\infty} P\{\xi_k = 1\} = 1 \ .$$

(b) Suppose (more strongly) that $\theta_k \downarrow 0$ and there exists a sequence $\{\epsilon_k\}$ with $0 < \epsilon_k < 1$ and $\epsilon_k \to 1$ such that

$$\sum_{n=k^*\epsilon_k}^{k} \theta_n^a + \frac{\gamma}{\delta} \log \theta_k \to \infty \quad \text{as } k\to\infty \ , \tag{2.17}$$

$$\sup_k \frac{\theta_{k^*\epsilon_k}}{\theta_k} < \infty \ . \tag{2.18}$$

Then

$$P\{\xi_k = 1\} = 1 + O(\theta_k^\gamma) \quad \text{as} \quad k \to \infty .$$

The proof of Theorem 2.3 will require the following lemmas.

Lemma 2.1 Let $\{s_k\}$ be a sequence of positive numbers with $s_k \to 0$ and

$$\sum_{k=1}^{\infty} s_k = \infty .$$

Then

$$\sum_{k=1}^{\infty} s_k \prod_{n=1}^{k-1} (1 - s_n) < \infty .$$

Proof Let

$$S_k = \sum_{n=1}^{k} s_k .$$

Now since $s_k \to 0$ and $S_k \to \infty$ we have

$$\exp(- S_{k-1}) = \exp(s_k) \exp(- S_k) \leq \frac{c}{S_k^2}$$

for some constant c. Hence

$$\begin{aligned} \sum_{k=1}^{\infty} s_k \prod_{n=1}^{k-1} (1 - s_n) &\leq \sum_{k=1}^{\infty} s_k \exp(- S_{k-1}) \\ &\leq c \cdot \sum_{k=1}^{\infty} \frac{s_k}{S_k^2} \\ &< \infty \end{aligned}$$

where the convergence of the last series follows from the Abel-Dini Theorem [20, p. 290]. □

Lemma 2.2 Let $b > a > 0$ and assume that $\theta_k \downarrow 0$ and

$$\sum_{k=1}^{\infty} \theta_k^a = \infty . \tag{2.19}$$

Then

$$\lim_{k\to\infty} \sum_{m=1}^{k} \theta_m^b \prod_{n=m+1}^{k} (1-\theta_n^a) = 0 .$$

Proof Let

$$p_k = \sum_{m=1}^{k} \theta_m^b \prod_{n=m+1}^{k} (1-\theta_n^a) .$$

Let $s_k = \theta_k^a$. Then for $K \in \mathbb{N}$

$$\begin{aligned} p_k &= \sum_{m=1}^{k} s_m^{b/a} \prod_{n=m+1}^{k} (1-s_n) \\ &\le K \cdot s_1^{b/a} \prod_{n=K}^{k} (1-s_n) + \theta_{K+1}^{\gamma} \sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1-s_n) \quad \forall\, k \ge K , \end{aligned}$$

where $\gamma = b - a > 0$. Hence

$$\limsup_{k\to\infty} p_k \le \theta_{K+1}^{\gamma} \sup_k \sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1-s_n) \tag{2.20}$$

since

$$\prod_{n=K}^{\infty} (1-s_n) = \prod_{n=K}^{\infty} (1-\theta_n^a) = 0$$

which follows from (2.19). Now

$$\sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1-s_n) = \sum_{m=1}^{k} s_m \prod_{n=1}^{m-1} (1-s_n)$$

which is established by induction on k. Hence by Lemma 2.1

$$\sup_k \sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1-s_n) < \infty . \tag{2.21}$$

Combining (2.20), (2.21) and letting $K\to\infty$ (so that $\theta_{K+1}^{\gamma}\to 0$) gives $p_k \to 0$ as required. □

Lemma 2.3 Let $b > a > 0$, $\gamma = b - a$, and assume that $\theta_k \downarrow 0$ and there exists a sequence $\{\epsilon_k\}$ with $0 < \epsilon_k < 1$ and $\epsilon_k \to 0$ such that

$$\sup_k \frac{\theta_{k\cdot\epsilon_k}}{\theta_k} < \infty . \tag{2.22}$$

Then

$$\sum_{m=k^{\bullet}\epsilon_k}^{k} \theta_m^b \prod_{n=m+1}^{k} (1-\theta_m^a) = O(\theta_k^\gamma) \quad \text{as} \quad k \longrightarrow \infty$$

Proof Let

$$p_k = \sum_{m=k^{\bullet}\epsilon_k}^{k} \theta_m^b \prod_{n=m+1}^{k} (1-\theta_n^a) .$$

Let $s_k = \theta_k^a$. Then

$$p_k = \sum_{m=k^{\bullet}\epsilon_k}^{k} s_m^{b/a} \prod_{n=m+1}^{k} (1-s_n)$$

$$\leq \theta_{k^{\bullet}\epsilon_k}^{\gamma} \sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1-s_n) . \tag{2.23}$$

Now

$$\sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1-s_n) = \sum_{m=1}^{k} s_m \prod_{n=1}^{m-1} (1-s_n)$$

which is established by induction on k. Hence by Lemma 2.1 there exists a constant c_1 such that

$$\sum_{m=1}^{k} s_m \prod_{n=m+1}^{k} (1-s_n) \leq c_1 . \tag{2.24}$$

Also from (2.22) there exists a constant c_2 such that

$$\theta_{k^{\bullet}\epsilon_k}^{\gamma} \leq c_2 \cdot \theta_k^{\gamma} . \tag{2.25}$$

Combining (2.23)-(2.25) gives $p_k = O(\theta_k^\gamma)$ as required. □

Proof of Theorem 2.3

(a) Define the events

$$C_{m,n} = \bigcap_{k=m}^{n} \{\xi_k \in \{2,3\}\} \qquad \forall\, n \geq m \ , \tag{2.26}$$

$$D_{m,n} = \{\xi_m = 1\} \cap C_{m+1,n} \qquad \forall\, n > m . \tag{2.27}$$

Then

$$\{\xi_k \in \{2,3\}\} = C_{1,k} \cup \bigcup_{m=1}^{k-1} D_{m,k}$$

and

$$P\{\xi_k \in \{2,3\}\} = PC_{1,k} + \sum_{m=1}^{k-1} PD_{m,k} \,. \tag{2.28}$$

Now using the lower bound in (2.15) and the Markov property, for $i \in \{2,3\}$

$$\begin{aligned} P\{C_{m,k} | \xi_m = i\} &\le P\{\xi_m = i\} \cdot \prod_{n=m}^{k-1} \max_{j=2,3} P\{\xi_{n+1} = 1 | \xi_n = j\} \\ &\le \prod_{n=m}^{k-1} \left(1 - \min_{j=2,3} P\{\xi_{n+1} \in \{2,3\} | \xi_n = j\}\right) \\ &\le \prod_{n=m}^{k-1} \left(1 - \min_{j=2,3} A_{j1} \theta_n^{\alpha_{j1}}\right) \\ &\le c_1 \cdot \prod_{n=m}^{k-1} \left(1 - \delta\, \theta_n^a\right) \qquad \forall\, k > m \,, \end{aligned} \tag{2.29}$$

for some constant c_1. Also, using the upper bound in (2.15), the Markov property, and (2.29)

$$\begin{aligned} PD_{m,k} &= \sum_{i=2,3} P\{\xi_m = 1\}\, p_{1i}^{(m,m+1)}\, P\{C_{m+1,k} | \xi_{m+1} = i\} \\ &\le 2 \cdot \max_{i=2,3} B_{1i}\, \theta_m^{\beta_{1i}} \cdot c_1 \prod_{n=m+1}^{k-1} (1 - \delta\, \theta_n^a) \\ &\le c_2 \cdot \theta_m^b \prod_{n=m+1}^{k-1} (1 - \delta\, \theta_n^a) \qquad \forall\, k > m \,, \end{aligned} \tag{2.30}$$

for some constant c_2. Hence from (2.29) and (2.16)

$$\begin{aligned} \lim_{k \to \infty} PC_{1,k} &\le c_1 \cdot \prod_{n=1}^{\infty} (1 - \delta\, \theta_n^a) \\ &= 0 \,, \end{aligned} \tag{2.31}$$

and from (2.30) and Lemma 2.2

$$\begin{aligned} \lim_{k \to \infty} \sum_{m=1}^{k-1} PD_{m,k} &\le \lim_{k \to \infty} c_2 \cdot \sum_{m=1}^{k-1} \theta_m^b \prod_{n=m+1}^{k-1} (1 - \delta\, \theta_n^a) \\ &= 0 \,. \end{aligned} \tag{2.32}$$

Combining (2.28), (2.31), and (2.32) gives $P\{\xi_k = 1\} \longrightarrow 1$ as required.

(b) Define $C_{m,n}$, $D_{m,n}$ as in (2.26), (2.27). Then

$$\{\xi_k \in \{2,3\}\} = C_{k^*\epsilon_k,k} \cup \bigcup_{m=k^*\epsilon_k}^{k-1} D_{m,k}$$

and

$$P\{\xi_k \in \{2,3\}\} = PC_{k^*\epsilon_k,k} + \sum_{m=k^*\epsilon_k}^{k-1} PD_{m,k} \,. \tag{2.33}$$

From (2.29) we have

$$\begin{aligned} PC_{k^*\epsilon_k,k} &\le c_1 \prod_{n=k^*\epsilon_k}^{k-1} (1 - \delta\, \theta_n^a) \\ &\le c_1 \exp\left(- \sum_{n=k^*\epsilon_k}^{k-1} \delta\, \theta_n^a \right) \\ &= c_1 \exp(\delta \theta_k^a) \exp\left[- \delta \left(\sum_{n=k^*\epsilon_k}^{k} \theta_n^a + \frac{\gamma}{\delta} \log \theta_k \right) \right] \theta_k^\gamma \\ &= o(\theta_k^\gamma) \qquad \text{as } k \to \infty \,, \end{aligned} \tag{2.34}$$

where the last equality follows from $\theta_k^a \to 0$ and (2.17). From (2.30) and Lemma 2.2

$$\begin{aligned} \sum_{m=k^*\epsilon_k}^{k-1} PD_{m,k} &\le c_2 \sum_{m=k^*\epsilon_k}^{k-1} \theta_m^b \prod_{n=m+1}^{k-1} (1 - \delta\, \theta_n^a) \\ &= O(\theta_k^\gamma) \qquad \text{as } k \to \infty \,. \end{aligned} \tag{2.35}$$

Combining (2.33)-(2.35) gives $P\{\xi_k = 1\} = 1 + O(\theta_k^\gamma)$ as required. □

The following corollary considers a choice of $\{\theta_k\}$ which will be seen to correspond to a temperature schedule $T_k = c/\log k$ for the annealing algorithm.

Corollary 2.1 Let Σ, a, b, γ, and δ be given as in Theorem 2.3. Assume that

$$\theta_k = \frac{1}{k^{1/c}}$$

where c is a positive constant.

(a) If $c \geq a$ then

$$\lim_{k\to\infty} P\{\xi_k = 1\} = 1 .$$

(b) If $c > a$ then

$$P\{\xi_k = 1\} = 1 + O(\theta_k^\gamma) \qquad \text{as } k\to\infty .$$

(c) If $c = a$ then

$$P\{\xi_k = 1\} = \begin{cases} 1 + O(\theta_k^\gamma) & \text{if } \gamma < \delta \\ 1 + O(\theta_k^\gamma \log k) & \text{if } \gamma = \delta \\ 1 + O(\theta_k^\delta) & \text{if } \gamma > \delta , \end{cases} \qquad \text{as } k\to\infty .$$

Proof We shall assume that $c = 1$; the general case follows easily.

(a) If $a \leq 1$ then

$$\sum_{k=1}^{\infty} \theta_k^a = \sum_{k=1}^{\infty} \frac{1}{k^a} = \infty$$

and Theorem 2.3(a) applies.

(b) Suppose $a < 1$. To apply Theorem 2.3(b) we must construct a sequence $\{\epsilon_k\}$ with $0 < \epsilon_k < 1$ and $\epsilon_k \to 1$ such that conditions (2.17), (2.18) are satisfied. Fix $0 < \eta < 1-a$ and let

$$\epsilon_k = 1 - \frac{1}{k^\eta} \qquad (k \text{ large}) .$$

Then for sufficiently large k

$$\sum_{n=k^*\epsilon_k}^{k} \theta_n^a = \sum_{n=k(1-k^{-\eta})}^{k} \frac{1}{n^a}$$

$$\geq \int_{k(1-k^{-\eta})}^{k} \frac{1}{x^a}\, dx$$

$$\geq \eta k^{1-a-\eta}$$

after evaluating the integrel and applying the Mean Value Theorem. Hence

$$\sum_{n=k^*\epsilon_k}^{k} \theta_n^a + \frac{\gamma}{\delta} \log \theta_k \geq \eta k^{1-a-\eta} - \frac{\gamma}{\delta} \log k \longrightarrow \infty \quad \text{as} \quad k \longrightarrow \infty ,$$

and consequently (2.17) is satisfied. (2.18) is also satisfied. Hence Theorem 2.3 (b) applies.

(c) Suppose $a = 1$. It is not apparent in this case how to construct the $\{\epsilon_k\}$ sequence which is necessary to apply Theorem 2.3 (b). However, we can directly use (2.28)-(2.30) to get the desired estimate of $P\{\xi_k = 1\}$. So, from (2.28)

$$P\{\xi_k \in \{2,3,\}\} = PC_{1,k} + \sum_{m=1}^{k-1} PD_{m,k} . \tag{2.36}$$

Now from (2.29)

$$PC_{1,k} \leq c_1 \prod_{n=1}^{k-1} (1 - \delta\, \theta_n^a)$$

$$\leq c_1 \exp\left(- \delta \sum_{n=1}^{k-1} \frac{1}{n} \right)$$

$$\leq c_1 \exp\left(- \delta \int_1^k \frac{1}{x}\, dx \right)$$

$$= \frac{c_1}{k^\delta} . \tag{2.37}$$

Also, from (2.30)

$$\sum_{m=1}^{k-1} PD_{m,k} \leq c_2 \sum_{m=1}^{k-1} \theta_m^b \prod_{n=m+1}^{k-1} (1 - \delta\, \theta_n^a)$$

$$\leq c_2 \sum_{m=1}^{k-1} \frac{1}{m^b} \exp\left(- \delta \sum_{n=m+1}^{k-1} \frac{1}{n} \right)$$

$$\leq c_2 \sum_{m=1}^{k-1} \frac{1}{m^b} \exp\left(- \delta \int_{m+1}^{k} \frac{1}{x}\, dx \right)$$

$$= \frac{c_2}{k^\delta} \sum_{m=1}^{k-1} \frac{1}{m^b} \cdot (m+1)^\delta$$

$$\leq \frac{2c_2}{k^\delta} \sum_{m=1}^{k-1} \frac{1}{m^{b-\delta}}$$

since $(p+q)^r \leq p^r + q^r$ for $p,q \geq 0$, $0 \leq r \leq 1$. Since $\delta \leq 1$ (use $\theta_1 = 1$ in (2.15)) and $b > a = 1$ we have $b - \delta > 0$. Hence

$$\sum_{m=1}^{k-1} PD_{m,k} \leq \frac{2c_2}{k^\delta} \left(1 + \int_1^k \frac{1}{x^{b-\delta}}\, dx \right)$$

$$= \begin{cases} c_3 \cdot \dfrac{1}{k^\delta} + c_4 \cdot \dfrac{1}{k^\gamma} & \text{if } \gamma \neq \delta \\ 2c_2\,(1 + \log k) \cdot \dfrac{1}{k^\gamma} & \text{if } \gamma = \delta \end{cases} \tag{2.38}$$

where c_3, c_4 are suitable constants. Combining (2.36)-(2.38) completes the proof of part (c). □

2.2.2 Sample Path Analysis

We now analyze the sample path behavior of a Markov chain $\{\xi_k\}$ with state space Σ which satisfies (2.13) and/or (2.14). We shall give (different) conditions such that

- $\{\xi_k\}$ visits a subset of Σ (infinitely often) with probability one
- $\{\xi_k\}$ visits a subset of Σ with probability less then one
- $\{\xi_k\}$ converges to (i.e. eventually stays in) a subset of Σ with probability one

It will be convenient to use the following notation. For J a subset of Σ define the events

$$\{\xi_k \in J \text{ i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{k>n} \{\xi_k \in J\} \ ,$$

$$\{\xi_k \in J \text{ a.a.}\} = \bigcup_{n=1}^{\infty} \bigcap_{k>n} \{\xi_k \in J\}$$

(i.o. and a.a. stand for *infinitely often* and *almost always*, respectively).

Our first theorem gives sufficient conditions under which $\{\xi_k\}$ visits a subset of Σ infinitely often with probability one.

Theorem 2.4 Assume that (2.13) holds. Let J be a subset of Σ and

$$a = \max_{i \in \Sigma \setminus J} \min_{j \in J} \alpha_{ij} < \infty \ . \tag{2.39}$$

Suppose

$$\sum_{k=1}^{\infty} \theta_k^a = \infty \ . \tag{2.40}$$

Then $P\{\xi_k \in J \text{ i.o.}\} = 1$.

Proof Let $I = \Sigma \setminus J$. Using (2.13) and the Markov property

$$\begin{aligned} P \bigcap_{k=m}^{n} \{\xi_k \in I\} &\le P\{\xi_m \in I\} \prod_{k=m}^{n-1} \max_{i \in I} P\{\xi_{k+1} \in I | \xi_k = i\} \\ &\le \prod_{k=m}^{n-1} \left(1 - \min_{i \in I} P\{\xi_{k+1} \in J | \xi_k = i\}\right) \\ &\le \prod_{k=m}^{n-1} \left(1 - \min_{i \in I} \sum_{j \in J} A\, \theta_k^{\alpha_{ij}}\right) \\ &\le \prod_{k=m}^{n-1} (1 - A\, \theta_k^a) \qquad \forall\, n > m \ . \end{aligned}$$

Hence

$$P \bigcap_{k=m}^{\infty} \{\xi_k \in I\} \le \prod_{k=m}^{\infty} (1 - A\, \theta_k^a) = 0 \qquad \forall\, m \ ,$$

where the divergence of the infinite product follows from the divergence of the infinite sum (2.40), and the Theorem follows. □

The next theorem gives sufficient conditions that $\{\xi_k\}$ visits a subset of Σ with probability strictly less than one, at least starting from certain initial states. Let $P_i(\cdot) = P\{\cdot|\xi_1 = i\}$ for all $i\in\Sigma$.

Theorem 2.5 Assume that (2.14) holds. Let J be a subset of Σ and

$$b = \max_{K\supset J} \min_{i\in\Sigma\setminus K} \min_{j\in K} \beta_{ij} > 0 . \tag{2.41}$$

Suppose that $\theta_k \to 0$ and

$$\sum_{k=1}^{\infty} \theta_k^b < \infty . \tag{2.42}$$

Then there exists an $i\in\Sigma$ such that

$$P_i \bigcup_{k=1}^{\infty} \{\xi_k\in J\} < 1 .$$

Proof Let J^* be a subset of Σ containing J which obtains the maximum in (2.41) and let $I^* = \Sigma\setminus J^*$. Let $\ell\in I^*$. Using (2.14) and the Markov property

$$\begin{aligned} P_\ell \bigcap_{k=1}^{n} \{\xi_k\in I^*\} &\geq \prod_{k=1}^{n-1} \min_{i\in I^*} P\{\xi_{k+1}\in I^*|\xi_k = i\} \\ &= \prod_{k=1}^{n-1} \left(1 - \max_{i\in I^*} P\{\xi_{k+1}\in J^*|\xi_k = i\}\right) \\ &\geq \prod_{k=1}^{n-1} \left(1 - \max_{i\in I^*} \sum_{j\in J^*} B\,\theta_k^{\beta_{ij}}\right) \\ &\geq \prod_{k=1}^{n-1} (1 - B|J^*|\theta_k^b) \qquad \forall n . \end{aligned}$$

Hence

$$P_\ell \bigcap_{k=1}^{\infty} \{\xi_k\in I^*\} \geq \prod_{k=1}^{\infty} (1 - B|J^*|\theta_k^b) > 0$$

where the convergence of the infinite product follows from the convergence of the infinite series (2.42), and the Theorem follows. □

Finally, we give a theorem which gives conditions such that $\{\xi_k\}$ converges to a subset of Σ with probability one, provided it visits that subset infinitely often with probability one.

Theorem 2.6 Assume (2.14) holds. Let J be a subset of Σ and

$$c = \min_{j\in J} \min_{i\in\Sigma\setminus J} \mathcal{J}_{ji} . \tag{2.43}$$

Suppose $\theta_k \to 0$ and

$$\sum_{k=1}^{\infty} \theta_k^c < \infty . \tag{2.44}$$

Under these conditions, if $P\{\xi_k\in J \text{ i.o.}\} = 1$ then $P\{\xi_k\in J \text{ a.a.}\} = 1$.

Proof Let $I = \Sigma\setminus J$. Using (2.14) and the Markov property

$$P\{\xi_k\in J, \xi_{k+1}\in I\} \le P\{\xi_k\in J\} \max_{j\in J} P\{\xi_{k+1}\in I|\xi_k = j\}$$

$$\le \max_{j\in J} \sum_{i\in I} B\, \theta_k^{\mathcal{J}_{ji}}$$

$$\le B|I|\, \theta_k^c .$$

Hence

$$\sum_{k=1}^{\infty} P\{\xi_k\in J, \xi_{k+1}\in I\} \le \sum_{k=1}^{\infty} B|I|\, \theta_k^c < \infty$$

by (2.44). Hence by the "first" Borel-Cantelli Lemma we must have $P\{\xi_k\in J, \xi_{k+1}\in I \text{ i.o.}\} = 0$, and it follows that $P\{\xi_k\in J \text{ a.a}\} = 1$ whenever $P\{\xi_k\in J \text{ i.o.}\} = 1$. □

2.3 Convergence of the Annealing Algorithm

In this Section we apply the results of 2.2 to obtain asymptotic properties of the annealing algorithm. Throughout this Section (2.3) we use the notation introduced in 2.1.

2.3.1 Bounds on the Transition Probabilities of the Annealing Chain

In order to apply the results in 2.2 we need to obtain bounds on the d-step transition probabilities $p_{ij}^{(k,k+d)}$ of the annealing chain $\{\xi_k\}$. Toward this end we make the following definitions. For every $i,j\in\Sigma$ and $d\in\mathbb{N}$ let $\Lambda_d(i,j)$ be the subset of Σ^{d+1} such that $(i = i_o,\ldots,i_d = j) \in \Lambda_d(i,j)$ if

$$p_{i_n i_{n+1}}^{(k,k+1)} > 0 \qquad \forall\, n = 0,\ldots,d-1 ,$$

for any $k\in\mathbb{N}$ (this definition is valid since $\{T_k\}$ is a positive sequence and so $p_{ij}^{(k,k+1)} > 0$ for all k whenever $p_{ij}^{(k,k+1)} > 0$ for some k). Hence $\Lambda_d(i,j)$ is just

the set of possible d-step transitions from state i to state j for the annealing chain. An alternate characterization of $\Lambda_d(i,j)$ is as follows: $(i = i_o,\ldots,i_d = j) \in \Lambda_d(i,j)$ iff for every $n = 0,\ldots,d-1$ *either*

(i) $q_{i_n i_{n+1}} > 0$ *or*

(ii) $i_{n+1} = i_n$ and $q(i_n,\ell) > 0$ for some $\ell \in \Sigma$ with $U(\ell) > U(i_n)$.

This characterization follows easily from (2.2).

For each $d \in \mathbb{N}$ let

$$U_d(i_o,\ldots,i_d) = \sum_{n=0}^{d-1} \max\{0, U(i_{n+1}) - U(i_n)\} ,$$

for all $i_o,\ldots,i_d \in \Sigma$, and

$$V_d(i,j) = \inf_{\lambda \in \Lambda_d(i,j)} U_d(\lambda) , \tag{2.45}$$

$$V(i,j) = \inf_d V_d(i,j) \tag{2.46}$$

for all $i,j \in \Sigma$. Note that the infinum in (2.46) is obtained for $d \leq |\Sigma|$. Also note that

$$V(i,j) \leq V(i,\ell) + V(\ell,j) \qquad \forall\ i,j,\ell \in \Sigma . \tag{2.47}$$

We shall call $V_d(i,j)$ the d-*step transition energy from* i *to* j, and $V(i,j)$ the *transition energy from* i *to* j.

The following theorem gives upper and lower bounds on the d-step transition probabilities of the annealing chain in terms of the d-step transition energy.

Theorem 2.7 Let $\{T_k\}$ be monotone nonincreasing and $d \in \mathbb{N}$. Then there exists positive numbers A, B such that

$$A \exp\left[-\frac{V_d(i,j)}{T_{k+d-1}}\right] \leq p_{ij}^{(k,k+d)} \leq B \exp\left[-\frac{V_d(i,j)}{T_k}\right] \qquad \forall\ i,j \in \Sigma . \tag{2.48}$$

Proof We prove the lower bound in (2.48); the upper bound is similar. Let

$$r_k(i,j) = \begin{cases} q_{ij} & \text{if } j \neq i \\ p_{ii}^{(k,k+1)} & \text{if } j = i \end{cases} \tag{2.49}$$

for all $i,j \in \Sigma$, and

$$\tilde{r}_k(i_o,\ldots,i_d) = \prod_{n=0}^{d-1} r_k(i_n,i_{n+1}) , \tag{2.50}$$

$$\tilde{r}(i_o,\ldots,i_d) = \inf_k \tilde{r}_k(i_o,\ldots,i_d) , \tag{2.51}$$

for all $i_o,\ldots,i_d \in \Sigma$. If $\lambda \in \Sigma^{d+1}$ then since $\{T_k\}$ is nonincreasing $\{\tilde{r}_k(\lambda)\}$ is nondecreasing and so $\tilde{r}(\lambda) = \tilde{r}_1(\lambda)$ obtains the infimum. Note that $\tilde{r}(\lambda) > 0$ for all $\lambda \in \Lambda_d(i,j)$, $i,j \in \Sigma$.

Now from (2.2) and (2.49)-(2.51) we have that

$$p_{ij}^{(k,k+d)} \geq \sum_{\lambda \in \Lambda_d(i,j)} \tilde{r}(\lambda) \exp\left[-\frac{U_d(\lambda)}{T_{k+d-1}}\right] \qquad \forall\, i,j \in \Sigma . \tag{2.52}$$

For each $i,j \in \Sigma$ if $V_d(i,j) < \infty$ let

$$N(i,j) = \{\lambda \in \Lambda_d(i,j) : \; U_d(\lambda) = V_d(i,j)\} \neq \varnothing$$

and set

$$a_{ij} = \sum_{\lambda \in N(i,j)} \tilde{r}(\lambda) > 0 ;$$

if $V_d(i,j) = \infty$ set $a_{ij} = 1$. Then from (2.52)

$$p_{ij}^{(k,k+d)} \geq A \exp\left[-\frac{V_d(i,j)}{T_{k+d-1}}\right] \qquad \forall\, i,j \in \Sigma$$

where $A = \min_{i,j \in \Sigma} a_{ij} > 0$. □

Remark We note that the proof of Theorem 2.7 is quite trivial, and we would like to point out that our reason for presenting it in detail is for comparison with the (more difficult) proof of the general state analog (Theorem 3.3) to come.

2.3.2 Convergence in Probability and Rate of Convergence for Two Local Minima

We now apply the results of 2.2.1 to establish the convergence in probability and rate of convergence for an annealing chain $\{\xi_k\}$ with an energy function $U(\cdot)$ with two local minima. We shall consider the following example in detail:

$$
\begin{aligned}
\text{(H)} \qquad & \Sigma = \{1,2,3\} \\
& U(1) < U(3) < U(2) \\
& q_{12},\, q_{21},\, q_{23},\, q_{32} > 0 \\
& q_{ij} = 0 \qquad \text{otherwise} .
\end{aligned}
$$

The annealing chain corresponding to (H) is illustrated by the transition diagram in Figure 2.1. Let

$$
\begin{aligned}
a &= U(2) - U(3) , \\
b &= U(2) - U(1) , \\
\gamma &= U(3) - U(1) , \\
\delta &= q_{32} \cdot q_{21} .
\end{aligned}
$$

Here is our theorem.

Theorem 2.8 Assume the conditions in (H).

(a) Suppose $T_k \downarrow 0$ and

$$
\sum_{k=1}^{\infty} \exp\left(-\frac{a}{T_k}\right) = \infty . \tag{2.53}
$$

Then

$$
\lim_{k \to \infty} P\{\xi_k = 1\} = 1 .
$$

(b) Suppose (more strongly) that $T_k \downarrow 0$ and there exists a sequence $\{\epsilon_k\}$ with $0 < \epsilon_k < 1$ and $\epsilon_k \to 1$ such that

$$
\sum_{n=k^{*}\epsilon_k}^{k} \exp\left(-\frac{a}{T_{2k}}\right) - \frac{\gamma}{\delta} \cdot \frac{1}{T_{2k}} \to \infty \qquad \text{as } k \to \infty , \tag{2.54}
$$

$$
\sup_k \left(\frac{1}{T_{2k}} - \frac{1}{T_{2k^{*}\epsilon_k}}\right) < \infty . \tag{2.55}
$$

Then

$$
P\{\xi_k = 1\} = 1 + O\left(\exp\left(-\frac{\gamma}{T_k}\right)\right) \qquad \text{as } k \to \infty . \tag{2.56}
$$

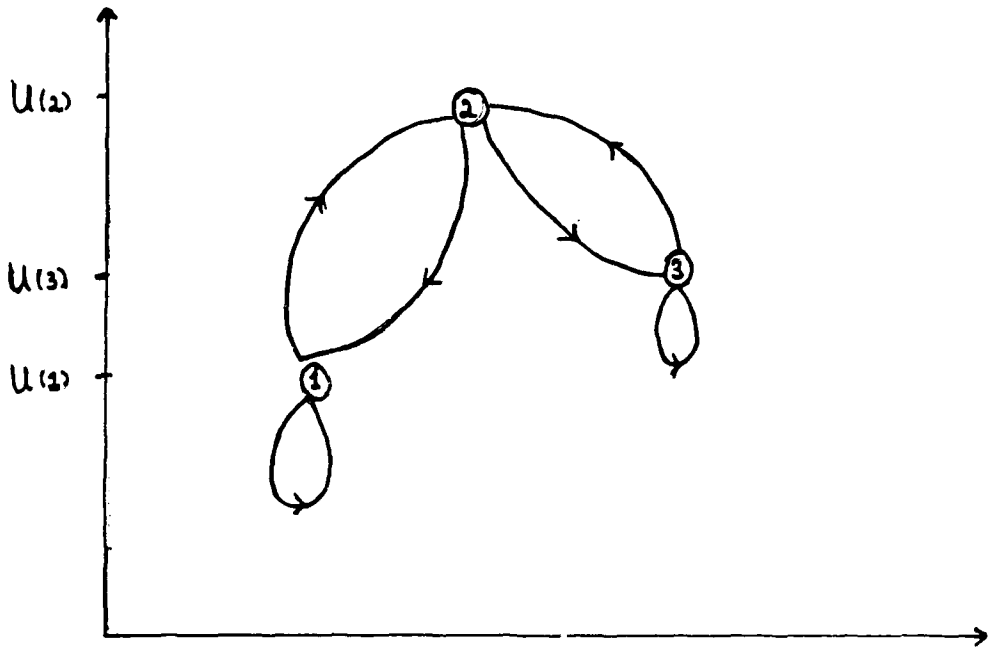


Figure 2.1. Transition Diagram for Annealing Chain with Two Local Minima

Proof Let

$$\eta_k = \xi_{2k}\,,\quad \varsigma_k = \xi_{2k+1}\,,$$
$$\theta_k = \exp\left(-\frac{1}{T_{2k}}\right),\quad \tau_k = \exp\left(-\frac{1}{T_{2k+1}}\right).$$

Then $\{\eta_k\}$, $\{\varsigma_k\}$ are Markov chains with 1-step transition matrices $\{R^{(k,k+1)}\} = \{[r_{ij}^{(k,k+1)}]\}$, $\{S^{(k,k+1)}\} = \{[s_{ij}^{(k,k+1)}]\}$, respectively, which satisfy

$$A_{ij}\theta_k^{\alpha_{ij}} < r_{ij}^{(k,k+1)} \le B_{ij}\theta_k^{\beta_{ij}}\,,$$

$$A_{ij}\tau_k^{\alpha_{ij}} \le s_{ij}^{(k,k+1)} \le B_{ij}\tau_k^{\beta_{ij}}\,,$$

for appropriate α_{ij}, β_{ij}, A_{ij}, B_{ij}, and it is clear that these constants may be chosen such that

$$a = \max\{\alpha_{21},\, \alpha_{31}\} < \infty\,,$$
$$b = \min\{\beta_{12},\, \beta_{13}\} > a\,,$$
$$\gamma = b - a\,,$$
$$\delta = A_{31}\,.$$

Hence we are (almost) in a position to apply Theorem 2.3 to $\{\eta_k\}$ and $\{\varsigma_k\}$.

Suppose that $T_k \downarrow 0$ and (2.53) holds. Since $\{T_k\}$ is nonincreasing, the divergence of the series in (2.53) implies that

$$\sum_{k=1}^{\infty} \theta_k^a = \infty\,,\quad \sum_{k=1}^{\infty} \tau_k^a = \infty\,.$$

Hence we may apply Theorem 2.3 (a) to $\{\eta_k\}$, $\{\varsigma_k\}$ to get

$$\lim_{k\to\infty} P\{\xi_{2k} = 1\} = \lim_{k\to\infty} P\{\eta_k = 1\} = 1\,,$$

$$\lim_{k\to\infty} P\{\xi_{2k+1} = 1\} = \lim_{k\to\infty} P\{\varsigma_k = 1\} = 1\,,$$

and hence

$$\lim_{k\to\infty} P\{\xi_k = 1\} = 1$$

which proves (a).

Suppose that $T_k \downarrow 0$ and (2.54), (2.55) hold. Now (2.54), (2.55) are equivalent to, respectively,

$$\sum_{n=k^*\epsilon_k}^{k} \theta_n^a + \frac{\gamma}{\delta} \log \theta_k \to \infty \qquad \text{as} \quad k \to \infty , \tag{2.57}$$

$$\sup_k \frac{\theta_{k^*\epsilon_k}}{\theta_k} < \infty . \tag{2.58}$$

Hence we may apply Theorem 2.3 (b) to $\{\eta_k\}$ to get

$$P\{\xi_{2k} = 1\} = P\{\eta_k = 1\} = 1 + O\left(\exp\left(-\frac{a}{T_{2k}}\right)\right) \qquad \text{as} \quad k \to \infty . \tag{2.59}$$

We make the following

Claim

$$\sum_{n=k^*\epsilon_k}^{k} \tau_n^a + \frac{\gamma}{\delta} \log \tau_k \to \infty \qquad \text{as} \quad k \to \infty , \tag{2.60}$$

$$\sup_k \frac{\tau_{k^*\epsilon_k}}{\tau_k} < \infty . \tag{2.61}$$

Suppose the Claim is true. Then we may apply Theorem 2.3 (b) to $\{\zeta_k\}$ to get

$$P\{\xi_{2k+1} = 1\} = P\{\zeta_k = 1\} = 1 + O\left(\exp\left(\frac{a}{T_{2k+1}}\right)\right) \qquad \text{as} \quad k \to \infty , \tag{2.62}$$

and it would follow from (2.59) and (2.62) that

$$P\{\xi_k = 1\} = 1 + O\left(\exp\left(-\frac{a}{T_k}\right)\right) \qquad \text{as} \quad k \to \infty ,$$

which would prove (b). It remains to prove the Claim.

Proof of Claim We first show that

$$\sup_k \left(\frac{1}{T_{2k+1}} - \frac{1}{T_{2k}}\right) < \infty . \tag{2.63}$$

Now

$$\frac{1}{T_{2k+1}} - \frac{1}{T_{2k}} < \frac{1}{T_{2(k+1)}} - \frac{1}{T_{2(k+1)^*\epsilon_k}} + \frac{1}{T_{2(k+1)^*\epsilon_k}} - \frac{1}{T_{2k}} .$$

In view of (2.55) it is enough to show

$$\sup_k \left(\frac{1}{T_{2(k+1)\epsilon_k}} - \frac{1}{T_{2k}} \right) < \infty ,$$

or since $\{T_k\}$ is nonincreasing,

$$2(k+1)\cdot\epsilon_k \leq 2k \qquad (k \text{ large}) .$$

Suppose this last inequality is not satisfied. Then there exists a sequence $\{k_n\}$ of positive integers with $k_n \uparrow \infty$ and

$$k_n \epsilon_{k_n} > k_n - \epsilon_{k_n} > k_n - 1 .$$

Hence

$$\begin{aligned} \liminf_{k\to\infty} & \left(\sum_{n=k\cdot\epsilon_k}^{k} \theta_n^a + \frac{\gamma}{\delta} \log \theta_k \right) \\ & \leq \lim_{n\to\infty} \left(\theta_{k_n}^a + \frac{\gamma}{\delta} \log \theta_{k_n} \right) \\ & = -\infty \end{aligned}$$

which contradicts (2.57). Hence (2.63) must be true. Now using (2.63) we obtain

$$\sup_k \left(\sum_{n=k\cdot\epsilon_k}^{k} \theta_n^a - \sum_{n=k\cdot\epsilon_k}^{k} \tau_n^a \right) < \sup_k \left(\theta_{k\cdot\epsilon_k} - \theta_{k+1} \right) < \infty ,$$

$$\sup_k (\log \theta_k - \log \tau_k) = \sup_k \left(\frac{1}{T_{2k+1}} - \frac{1}{T_{2k}} \right) < \infty ,$$

$$\sup_k \left(\frac{\tau_{k\cdot\epsilon_k}}{\tau_k} - \frac{\theta_{k\cdot\epsilon_k}}{\theta_k} \right) < \sup_k \exp \left(\frac{1}{T_{2k+1}} - \frac{1}{T_{2k}} \right) < \infty ,$$

and (2.60), (2.61) now follow from (2.57), (2.58). This completes the proof of the Claim and hence the Theorem. □

Corollary 2.2 Assume the conditions in (H). Let

$$T_k = \frac{c}{\log k} \qquad (k \text{ large})$$

where c is a positive constant.

(a) If $c \geq a$ then

$$\lim_{k \to \infty} P\{\xi_k = 1\} = 1 .$$

(b) If $c > a$ then

$$P\{\xi_k = 1\} = 1 + O\left(\exp\left(-\frac{\gamma}{T_k}\right)\right) \quad \text{as } k \to \infty . \tag{2.64}$$

(c) If $c = a$ then

$$P\{\xi_k = 1\} = \begin{cases} 1 + O\left(\exp\left(-\dfrac{\gamma}{T_k}\right)\right) & \text{if } \gamma < \bar{\delta} \\ 1 + O\left(\exp\left(-\dfrac{\gamma}{T_k} + \log\log k\right)\right) & \text{if } \gamma = \bar{\delta} \\ 1 + O\left(\exp\left(-\dfrac{\bar{\delta}}{T_k}\right)\right) & \text{if } \gamma > \bar{\delta}, \text{ as } k \to \infty , \end{cases} \tag{2.65}$$

where $\bar{\delta} = \delta/2$.

Proof We may use Corollary 2.1 by appropriately identifying variables. Let

$$\eta_k = \xi_{2k} , \quad \zeta_k = \xi_{2k+1} ,$$

and

$$\theta_k = \frac{1}{k^{1/c}} .$$

Then $\{\eta_k\}$, $\{\zeta_k\}$ are Markov chains with one step transition matrices $\{R^{(k,k+1)}\} = \{[r_{ij}^{(k,k+1)}]\}$, $\{S^{(k,k+1)}\} = \{[s_{ij}^{(k,k+1)}]\}$, respectively, which satisfy

$$A_{ij}\, \theta_k^{\alpha_{ij}} \leq r_{ij}^{(k,k+1)}, s_{ij}^{(k,k+1)} \leq B_{ij}\, \theta_k^{\beta_{ij}} \qquad (k \text{ large})$$

for appropriate α_{ij}, β_{ij}, A_{ij}, B_{ij}, and these constants may be chosen such that

$$a = \max\{\alpha_{21},\alpha_{31}\} < \infty ,$$
$$b = \min\{\beta_{12},\beta_{13}\} > a ,$$
$$\gamma = b - a ,$$
$$\delta = A_{31} .$$

Hence we may apply Corollary 2.1 (a)-(c) to $\{\eta_k\}$, $\{\varsigma_k\}$ to get the corresponding (a)-(c) here. □

Remarks on Theorem 2.8 and Corollary 2.2

(1) Theorem 2.8 (a) is a simple case of Theorem 2.2 (Hajek's Theorem) since $a = \Delta(2) = \Delta^*$, the optimal constant (see (2.10)).

(2) We compare our results with the rate of convergence (2.9) given by Mitra et. al. First, Theorem 2.8 (b) gives the rate of convergence of $P\{\xi_k = 1\}$ to 1 *for* nonparametric temperature schedules, in particular schedules *not* of the form $T_k = c/\log k$. This is possible essentially due to the application of the Abel-Dini Theorem on infinite series in the proof of Lemma 2.1. (2.9) is valid only for temperature schedules of the form $T_k = c/\log k$. Second, Corollary 2.2 (b), (c) gives the rate of convergence for temperature schedules of the form $T_k = c/\log k$ for $c \geq a$, whereas (2.9) only holds for $c \geq r\,L = 2\cdot[U(2) - U(1)] > U(2) - U(3) = a$. Furthermore, for $c \geq r\,L$ where (2.9) does hold, (2.64) is in general tighter:

$$\exp\left(-\frac{\gamma}{T_k}\right) = \frac{1}{k^{\beta}} \leq \frac{1}{k^{\min\{\alpha,\beta\}}} .$$

Recall that Mitra et. al. suggest choosing $c \geq r\,L$ such that $\min\{\alpha,\beta\}$ is maximized (see (2.9)). Our results suggest choosing

$$c = \begin{cases} a & \text{if} \quad \gamma \leq \delta \\ a + \epsilon & \text{if} \quad \gamma > \delta \end{cases}$$

where $0 < \epsilon < a\ [(\gamma/\delta) - 1]$ (see (2.64) and (2.65)). We want to stress that (2.9) holds for general $U(\cdot)$ whereas we have not been able to extend Theorem 2.8 and Corollary 2.2 to a $U(\cdot)$ with more than two local minma.

(3) The proof of Theorem 2.8 and Corollary 2.2 (which rely on Theorem 2.3 and Corollary 2.1) show that there are two factors which limit the rate at which $P\{\xi_k = 1\}$ converges to 1. One factor corresponds to the rate at which the annealing chain makes transitions from state 1 to state 3 and back. For temperature schedules of the form $T_k = c/\log k$ this factor dominates for $c > a$ and has a characteristic time scale $1/\gamma$. Note that $\gamma = U(3) - U(1)$

depends only on the energy function U(·). The other factor corresponds to the rate at which the annealing chain makes it first transition from state 3 to state 1. For temperature schedules of the form $T_k = c/\log k$ this factor is only important for $c = a$ and has characteristic time scale $1/\delta$. Note that $\delta = q_{32}q_{21}/2$ does not depend on the energy function U(·). We wonder whether there is some physical significance to all of this.

2.3.3 Sample Path Analysis

We now apply the results of 2.2.2 to analyze the sample path behavior of the annealing chain $\{\xi_k\}$. To avoid trivialities we will need the following assumptions:

(P1) Every $i \in \Sigma \backslash S$ can reach some $j \in S$

(P2) There exists an $i \in \Sigma \backslash S$ such that for every $j \in S$, i can only reach j at height greater than U(i).

The following theorem gives conditions under which the annealing chain $\{\xi_k\}$ visits S infinitely often with probability one. Let

$$V^* = \max_{i \in \Sigma \backslash S} \min_{j \in S} V(i,j) \tag{2.66}$$

Note that (P1) holds iff $V^* < \infty$.

Theorem 2.9 Assume (P1). Let $\{T_k\}$ be monotone nonincreasing and

$$\sum_{k=1}^{\infty} \exp\left(-\frac{V^*}{T_k}\right) = \infty . \tag{2.67}$$

Then $P\{\xi_k \in S \text{ i.o.}\} = 1$.

Proof We first show there exists a $d \in \mathbb{N}$ such that

$$V^* = \max_{i \in \Sigma \backslash S} \min_{j \in S} V_d(i,j) . \tag{2.68}$$

For every $i \in \Sigma \backslash S$ there exists a $d_i \in \mathbb{N}$ such that

$$\min_{j \in S} V_{d_i}(i,j) = \min_{j \in S} V(i,j) \leq V^* .$$

Let $d^* = \max_{i \in \Sigma \backslash S} d_i$. Now it is easy to see that for every $i \in \Sigma$

$$\min_{j \in S} V_n(i,j) \leq \min_{j \in S} V_m(i,j) \qquad \forall\, n \geq m .$$

Hence for every $i \in \Sigma \backslash S$

$$\min_{j\in S} V_d(i,j) = \min_{n\le d^*} \min_{j\in S} V_n(i,j) \le V^*$$

and (2.68) follows by setting $d = d^*$.

Next, from Theorem 2.7 there exists a positive number A such that

$$p_{ij}^{(k,k+d)} \ge A \exp\left[-\frac{V_d(i,j)}{T_{k+d-1}}\right] \qquad \forall\, i,j\in\Sigma\,.$$

Let

$$\tilde{\xi}_k = \xi_{kd},\ \ \theta_k = \exp\left(-\frac{1}{T_{kd+d-1}}\right),$$

and

$$\alpha(i,j) = V_d(i,j) \qquad \forall\, i,j\in\Sigma\,. \tag{2.69}$$

Then $\{\tilde{\xi}_k\}$ is a Markov chain with 1-step transition matrices $\{\tilde{P}^{(k,k+1)}\} = \{[\tilde{p}_{ij}^{(k,k+1)}]\}$ which satisfy

$$\tilde{p}_{ij}^{(k,k+1)} \ge A\ \theta_k^{\alpha_{ij}} \qquad \forall\, i,j\in\Sigma\,.$$

Let

$$a = \max_{i\in\Sigma\setminus S} \min_{j\in S} \alpha_{ij}\,.$$

By (2.68) and (2.69) $a = V^*$. Hence since $\{T_k\}$ is nonincreasing the divergence of the series in (2.67) implies

$$\sum_{k=1}^{\infty} \theta_k^a = \infty\,.$$

Hence we may apply Theorem 2.4 to $\{\tilde{\xi}_k\}$ with $J = S$ to get $P\{\tilde{\xi}_k\in S \text{ i.o.}\} = 1$ and so $P\{\xi_k\in S \text{ i.o.}\} = 1$. □

Remark Clearly $V^* > \Delta^*$, the optimal constant (see (2.10), (2.66)). Hence (assuming strong irreducibility and weak reversibility) Theorem 2.2 is a much stronger result. However, the importance of Theorem 2.9 is that it can be extended to a general state version of the annealing algorithm under essentially the condition that the state space be a compact metric space and the energy function be continuous. This will be done in Chapter 3.

The next theorem gives conditions under which the annealing chain $\{\xi_k\}$ visits S with probability strictly less than one. Let

$$V_1 = \max_{K \supset S} \min_{i \in \Sigma \setminus K} \min_{j \in S} V(i,j) . \tag{2.70}$$

Note that (P2) and (2.47) imply $V_1 > 0$.

Theorem 2.10 Assume (P2). Let $T_k \to 0$ and

$$\sum_{k=1}^{\infty} \exp\left(-\frac{V_1}{T_k}\right) < \infty .$$

Then there exists an $i \in \Sigma$ such that

$$P_i \bigcup_{k=1}^{\infty} \{\xi_k \in S\} < 1 .$$

Proof From Theorem 2.7 there exists a positive number B such that

$$p_{ij}^{(k,k+1)} \leq B \exp\left[-\frac{V(i,j)}{T_k}\right] \qquad \forall\, i,j \in \Sigma .$$

Theorem 2.5 may be applied to $\{\xi_k\}$ in an obvious manner. □

Finally, we give a theorem which gives conditions such that the annealing chain $\{\xi_k\}$ converges to S with probability one, provided it visits S infinitely often with probability one. Let

$$V_2 = \min_{j \in S} \min_{i \in \Sigma \setminus S} V(j,i) . \tag{2.71}$$

Theorem 2.11 Let $T_k \to 0$ and

$$\sum_{k=1}^{\infty} \exp\left(-\frac{V_2}{T_k}\right) < \infty .$$

If $P\{\xi_k \in S \text{ i.o.}\} = 1$ then $P\{\xi_k \in S \text{ a.a.}\} = 1$.

Proof From Theorem 2.7 there exits a positive number B such that

$$p_{ij}^{(k,k+1)} \leq B \exp\left[-\frac{V(i,j)}{T_k}\right] \qquad \forall\, i,j \in \Sigma .$$

Theorem 2.6 may be applied to $\{\xi_k\}$ in an obvious manner. □

Remark Theorem 2.2 or 2.9 may be combined with Theorem 2.11 to give conditions under which the annealing chain $\{\xi_k\}$ converges to S with probability one. Note, however, that is is not always possible to do this since it is not in general true that $V_2 > V^*$ or even $V_2 > \Delta^*$ (see (2.10), (2.66), (2.71)).

2.4 Annealing Algorithm with Noisy Energy Measurements

In this Section we consider a modification of the annealing algorithm so as to allow for noisy measurements of the energy differences which are used in selecting successive states. This is important when the energy differences cannot be computed exactly or when it is simply too costly to do so. Using the notation introduced in 2.1 we construct the modified annealing chain as follows. At time k, given the current state is i we select a candidate state j with probability q_{ij}. We assume that the energy difference $U(j) - U(i)$ is measured with (additive) noise, which is independent of states and candidate states at times less than or equal to k, and noise at times less than k. The exponent of the energy difference plus noise is then used to determine whether a transition is made from i to j. More precisely, let $\{w_k\}$ be a sequence of $\mathbb{R}$-valued independent random variables. Construct a Σ-valued discrete-time process $\{\hat{\xi}_k\}$ with $\hat{\xi}_{k+1}$ conditionally independent of $\hat{\xi}_1,\ldots,\hat{\xi}_{k-1}$ and $w_1,\ldots,w_{k-1}$ given $\hat{\xi}_k$ and w_k, and

$$P\{\hat{\xi}_{k+1} = j | \hat{\xi}_k = i,\ w_k = w\}$$

$$= \begin{cases} q_{ij} \exp\left[-\dfrac{U(j) - U(i) + w}{T_k}\right] & \text{if } U(j) - U(i) + w > 0,\ j \neq i, \\ q_{ij} & \text{if } U(j) - U(i) + w \leq 0,\ j \neq i, \end{cases}$$

for all $i,j \in \Sigma$. It is easy to see that $\{\hat{\xi}_k\}$ is a Markov chain. Let $\{\hat{P}^{(k,k+1)}\} = \{[\hat{p}_{ij}^{(k,k+1)}]\}$ be the 1-step transition matrices for $\{\hat{\xi}_k\}$. Then since w_k is independent of $\hat{\xi}_k$ we have

$$\begin{aligned} \hat{p}_{ij}^{(k,k+1)} &= E\{P\{\hat{\xi}_{k+1} = j | \hat{\xi}_k,\ w_k\} | \hat{\xi}_k = i\} \\ &= E\{P\{\hat{\xi}_{k+1} = j | \hat{\xi}_k = i,\ w_k\}\} \\ &= \int_{\{w > U(i) - U(j)\}} q_{ij} \exp\left[-\frac{U(j) - U(i) + w}{T_k}\right] dF_k(w) \\ &\quad + q_{ij}\, P\{w_k \leq U(i) - U(j)\} \qquad \forall\, j \neq i, \end{aligned} \tag{2.72}$$

where $F_k(\cdot)$ is the distribution function for w_k. We shall call $\{\hat{\xi}_k\}$ the *annealing chain modified for noisy energy measurements.* In the sequel we

shall only consider the case where w_k is Gaussian with mean 0 and variance $\sigma_k^2 > 0$. Hence (2.72) can be written as

$$\hat{p}_{ij}^{(k,k+1)} = \int_{U(i) - U(j)}^{\infty} q_{ij} \exp\left[- \frac{U(j) - U(i) + w}{T_k} \right] dN(0,\sigma_k^2)\,(w)$$

$$+ q_{ij}\, N(0,\sigma_k^2)\,(-\infty,\, U(i) - U(j)] \qquad \forall\, j \neq i\,. \tag{2.73}$$

The following theorem shown that if the noise variance goes to zero fast enough then the 1-step transition probabilities for the annealing chain modified for (Gaussian additive) noisy measurements are asymptotically equivalent to the 1-step transition probabilities for the unmodified annealing chain.

Theorem 2.12 *If*

$$\sigma_k^2 = o(T_k^4) \qquad \text{as } k \to \infty$$

then

$$\hat{p}_{ij}^{(k,k+1)} \sim \begin{cases} q_{ij} \exp\left[- \dfrac{U(j) - U(i)}{T_k} \right] & \text{if } U(j) > U(i) \\ q_{ij} & \text{if } U(j) \leq U(i)\,,\ j \neq i\,, \end{cases} \tag{2.74}$$

as $k \to \infty$, for all $i,j \in \Sigma$.

Proof Fix $i,j \in \Sigma$ with $j \neq i$ and $q_{ij} > 0$. Let

$$a_k = \int_{U(i) - U(j)}^{\infty} q_{ij} \exp\left[- \frac{U(j) - U(i) + w}{T_k} \right] dN(0,\sigma_k^2)\,(w)$$

$$b_k = q_{ij}\, N(0,\sigma_k^2)\,(-\infty,\, U(i) - U(j)]$$

so that (2.73) becomes

$$\hat{p}_{ij}^{(k,k+1)} = a_k + b_k\,. \tag{2.75}$$

Clearly,

$$\lim_{k \to \infty} a_k = 0 \qquad \text{if } U(j) < U(i)\,, \tag{2.76}$$

$$\lim_{k\to\infty} b_k = \begin{cases} q_{ij} & \text{if } U(j) < U(i) \\ \dfrac{q_{ij}}{2} & \text{if } U(j) = U(i)\,. \end{cases} \tag{2.77}$$

We make the following

Claim

$$a_k \sim q_{ij} \exp\left[-\frac{U(j)-U(i)}{T_k}\right] \quad \text{if } U(j) > U(i) \text{ and } \sigma_k^2 = o(T_k^4) \tag{2.78}$$

$$a_k \longrightarrow \frac{q_{ij}}{2} \quad \text{if } U(j) = U(i) \text{ and } \sigma_k^2 = o(T_k^2) \tag{2.79}$$

$$b_k = o\left(\exp\left[-\frac{U(j)-U(i)}{T_k}\right]\right) \quad \text{if } U(j) > U(i) \text{ and } \sigma_k^2 = o(T_k) \tag{2.80}$$

as $k\to\infty$.

Suppose the Claim is true. Then combining (2.75)-(2.80) gives (2.74) if $\sigma_k^2 = o(T_k^4)$, as required. It remains to prove the Claim.

Proof of Claim We have

$$\begin{aligned} a_k &= \int_{U(i)-U(j)}^{\infty} q_{ij} \exp\left[-\frac{U(j)-U(i)+w}{T_k}\right] dN(0,\sigma_k^2)\,(w) \\ &= q_{ij} \exp\left[-\frac{U(j)-U(i)}{T_k}\right] \int_{T_k[U(i)-U(j)]}^{\infty} e^{-w}\, dN(0,\sigma_k^2/T_k^2)\,(w) \end{aligned} \tag{2.81}$$

after a change of variable. Choose $W < 0$ and let

$$f(w) = \begin{cases} e^{-w} & \text{if } w \geq W \\ e^{-W} & \text{if } w < W\,. \end{cases}$$

Then for sufficiently large k

$$\begin{aligned} &\int_{T_k[U(i)-U(j)]}^{\infty} e^{-w}\, dN(0,\sigma_k^2/T_k^2)\,(w) \\ &= \int_{-\infty}^{\infty} f(w)\, dN(0,\sigma_k^2/T_k^2)\,(w) - \int_{-\infty}^{T_k[U(i)-U(j)]} f(w)\, dN(0,\sigma_k^2/T_k^2)\,(w) \end{aligned} \tag{2.82}$$

We analyze these last two integrels as follows. First, if $\sigma_k^2 = o(T_k^2)$ then

$N(0,\sigma_k^2/T_k^2)(\cdot)$ converges weakly to the unit measure concentrated at the origin, and since $f(\cdot)$ is a bounded and continuous function on $\mathbb{R}$,

$$\lim_{k\to\infty} \int_{-\infty}^{\infty} f(w)\, dN(0,\sigma_k^2/T_k^2)(w) = f(0) = 1 . \tag{2.83}$$

Next, if $U(j) > U(i)$ and $\sigma_k^2 = o(T_k^4)$ then

$$\begin{aligned}
&\int_{-\infty}^{T_k[U(i)-U(j)]} f(w)\, dN(0,\sigma_k^2/T_k^2)(w) \\
&\le e^{-W} N(0,\sigma_k^2/T_k^2)(-\infty, T_k[U(i)-U(j)]] \\
&= e^{-W} N(0,1)((T_k^2/\sigma_k)\cdot[U(j)-U(i)], \infty) \\
&\le e^{-W} \exp\left[-\frac{[U(j)-U(i)]^2}{2(\sigma_k^2/T_k^4)}\right] \\
&\to 0 \quad \text{as } k\to\infty ,
\end{aligned} \tag{2.84}$$

where we have used $N(0,1)(x,\infty) \le \exp(-x^2/2)$ for $x \ge 0$. Combining (2.81)-(2.84) gives (2.78). (2.79) may be proved similarly by taking $W = 0$ above. As for (2.80), if $U(j) > U(i)$ and $\sigma_k^2 = o(T_k)$ then

$$\begin{aligned}
b_k &= N(0,\sigma_k^2)(-\infty, U(i)-U(j)] \\
&= N(0,1)((1/\sigma_k)\cdot[U(j)-U(i)],\infty) \\
&\le \exp\left[-\frac{[U(j)-U(i)]^2}{2\sigma_k^2}\right] \\
&= o\left(\exp\left[-\frac{U(j)-U(i)}{T_k}\right]\right) \quad \text{as } k\to\infty ,
\end{aligned}$$

again using $N(0,1)(x,\infty) \le \exp(-x^2/2)$ for $x \ge 0$. This completes the proof of the Claim and hence the Theorem. □

The asymptotic behavior of the annealing chain modified for (Gaussian additive) noisy measurements follows immediately:

Corollary 2.3 If

$$\sigma_k^2 = o(T_k^4) \qquad \text{as} \quad k \longrightarrow \infty$$

then Theorems 2.1, 2.2, 2.7-2.11 hold with $\{\xi_k\}$ by $\{\dot{\xi}_k\}$.

Remarks

(1) The Corollary is more or less obvious, since the convergence in (2.74) is uniform for $i,j \in \Sigma$ (since Σ is finite); we leave the details to the reader.

(2) We have reason to believe that $\sigma_k^2 = o(T_k^4)$ is quite conservative and that $\sigma_k^2 = o(T_k^2)$ may suffice.

CHAPTER III
GENERAL STATE ANNEALING TYPE ALGORITHMS

3.1 Introduction to the General State Annealing Algorithm

In Chapter 2 we discussed the annealing algorithm as introduced by Kirkpatrick [19] and Cerny [3] for combinatorial optimization. In this Section we extend the annealing algorithm for optimization on general spaces. The general state annealing algorithm will consist of simulating a nonstationary Markov chain whose state space is the domain of the cost function (called energy) to be minimized. This Markov chain will be a general state space analog of the finite state annealing chain described in Chapter 2. As far as we know, no one has given a careful formulation of such an algorithm and proved a convergence result. Indeed, there even seems to be some question regarding conditions under which the Metropolis algorithm, i.e., the annealing algorithm at a fixed temperature, may be used for sampling from a continuous Gibbs distribution (c.f. [16]). Geman and independently Grenander [13] have suggested using diffusions for optimization on multi-dimensional Euclidean space. This approach and its relationship to the general state annealing algorithm is described in Chapter 4.

We first give some standard general state space Markov chain notation (c.f. [6], [27]). Let Σ be an arbitrary set and let B be σ-field of subsets of Σ. $P(\cdot,\cdot)$ is a stochastic transition function on (Σ,B) if

- for every $A\in B$ $P(\cdot,A)$ is B-measurable
- for every $x\in\Sigma$ $P(x,\cdot)$ is a probability measure on (Σ,B).

$\{P_k(\cdot,\cdot)\}$ are the 1-step transition functions for a Markov chain $\{\xi_k\}$ with state space Σ if for every $k\in\mathbb{N}$ $P_k(\cdot,\cdot)$ is a stochastic transition function on (Σ,B) and

$$P\{\xi_{k+1}\in A|\xi_k\} = P_k(\xi_k,A) \qquad \text{w.p. } 1 \tag{3.1}$$

for all $A\in B$. Conversely given a sequence $\{P_k(\cdot,\cdot)\}$ of stochastic transition functions on (Σ,B) we can construct on a suitable probability space (Ω,F,P) a Markov chain $\{\xi_k\}$ with state space Σ which satisfies (3.1). For each $d\in\mathbb{N}$ let

$$P^{(k,k+d)}(x,A) = \int P_k(x,dx_1) \cdots \int P_{k+d-2}(x_{d-2},dx_{d-1})\, P_{k+d-1}(x_{d-1},A)$$

for all $x\in\Sigma$ and $A\in B$. $P^{(k,k+d)}(\cdot,\cdot)$ is a stochastic transition function on (Σ,B) and

$$P\{\xi_{k+d}\in A|\xi_k\} = P^{(k,k+d)}(\xi_k,A) \qquad \text{w.p. } 1$$

for all $A\in B$. It will be convenient to have a fixed version of the conditional probability of ξ_{k+d} given ξ_k which we define by

$$P\{\xi_{k+d}\in A|\xi_k = x\} = P^{(k,k+d)}(x,A)$$

for all $x\in\Sigma$ and $A\in B$.

It is characteristic of the theory of Markov chains with general state space that there exists an auxiliary σ-finite measure usually denoted by $\phi(\cdot)$, i.e., the state space is a σ-finite measure space (Σ,B,ϕ). We shall adopt this framework. We now define the general state annealing algorithm. Let $U(\cdot)$ be a nonnegative B-measurable function on Σ, which we shall call the *energy function*. The goal is to find a point in Σ which minimizes or nearly minimizes $U(\cdot)$. Let $\{T_k\}$ be a sequence of positive numbers, which we shall call the *temperature schedule*. Let $q(\cdot,\cdot)$ be a nonnegative $B\times B$-measurable function on $\Sigma\times\Sigma$ such that

$$\int q(x,y)\, \phi(dy) = 1 \qquad \forall\, x\in\Sigma\,.$$

Now let $\{\xi_k\}$ be the Markov chain with state space Σ and 1-step transition functions $\{P_k(\cdot,\cdot)\}$ given by

$$P_k(x,A) = \int_A q(x,y)\, s_k(x,y)\, \phi(dy) + \gamma_k(x)\, \delta(x,A) \tag{3.2}$$

for all $x\in\Sigma$ and $A\in B$, where

$$s_k(x,y) = \begin{cases} \exp\left[-\dfrac{U(y)-U(x)}{T_k}\right] & \text{if } U(y) > U(x) \\ 1 & \text{if } U(y) \le U(x)\ , \end{cases}$$

$$\gamma_k(x) = 1 - \int q(x,y)\, s_k(x,y)\, \phi(dy)\ ,$$

and $\delta(x,\cdot)$ is the unit measure concentrated at x, for all $x,y\in\Sigma$ (note that

Fubini's Theorem guarantees that $P_k(\cdot,\cdot)$ defined by (3.2) is a valid stochastic transition function). We shall denote by $p_k(x,\cdot)$ the density of the ϕ-absolutely continuous component of $P_k(x,\cdot)$ and by $p^{(k,k+d)}(x,\cdot)$ the density of the ϕ-absolutely continuous component of $P^{(k,k+d)}(x,\cdot)$. Note that if Σ is finite and $\phi(\cdot)$ is counting measure then $\{\xi_k\}$ is just the finite state annealing chain of Chapter 2 with $q_{ij} = q(i,j)$ (see (2.1)). Hence we shall also call $\{\xi_k\}$ the *annealing chain,* and the algorithm which simulates the sample paths of $\{\xi_k\}$ with $T_k \to 0$ the *annealing algorithm.*

The motivation behind the general state annealing algorithm is similar to the finite state case as described in Chapter 2. Let

$$Q(x,A) = \int_A q(x,y)\, \phi(dy)$$

for all $x \in \Sigma$ and $A \in \mathcal{B}$. $Q(\cdot,\cdot)$ is a stochastic transition function on $(\Sigma,\mathcal{B})$. For each $d \in \mathbb{N}$ let

$$Q^{(d)}(x,A) = \int Q(x,dx_1) \cdots \int Q(x_{d-2},dx_{d-1})\, Q(x_{d-1},A)$$

for all $x \in \Sigma$ and $A \in \mathcal{B}$. The following definitions generalize the familiar finite state definitions. We shall say that $Q(\cdot,\cdot)$ is *irreducible* if for every $x \in \Sigma$ and $A \in \mathcal{B}$ with $\phi(A) > 0$ there exists $d \in \mathbb{N}$ such that $Q^{(d)}(x,A) > 0$. We shall say that $Q(\cdot,\cdot)$ is *symmetric* if $q(x,y) = q(y,x)$ for all $x,y \in \Sigma$. Suppose $Q(\cdot,\cdot)$ is irreducible and symmetric, and let $\{\xi_k^T\}$ be the stationary chain with 1-step (stationary) transition function $P^T(\cdot,\cdot)$ given by the r.h.s. of (3.2) with $T_k = T$, a positive constant. Suppose that $0 < \phi(\Sigma) < \infty$. Then it can be shown that $P^T(\cdot,\cdot)$ has an invariant Gibbs measure $\Pi^T(\cdot)$, i.e.,

$$\Pi^T(A) = \int \Pi^T(dx)\, P^T(x,A) \qquad \forall\, A \in \mathcal{B}\ ,$$

where

$$\Pi^T(A) = \frac{\int_A \exp[-U(x)/T]\phi(dx)}{\int \exp[-U(y)/T]\phi(dy)} \qquad \forall\, A \in \mathcal{B}\ .$$

This follows from the detailed reversibility

$$\pi^T(x)p^T(x,y) = \pi^T(y)p^T(y,x)\ ,$$

valid for $\phi \times \phi$-a.e. $x, y \in \Sigma$, where $\pi^T(\cdot)$ and $p^T(x,\cdot)$ are the densities of the ϕ-absolutely continuous components of $\Pi^T(\cdot)$ and $P^T(x,\cdot)$, respectively. Furthermore, $Q(\cdot,\cdot)$ irreducible and symmetric implies that $\{\xi_k^T\}$ is an irreducible† (and aperiodic) chain and if a certain condition of Doeblin [6, p.

†A stationary chain is irreducible if its 1-step (stationary) transition function is irreducible.

192] is satisfied, then by a version of the Markov Convergence Theorem [6, p. 199]

$$\lim_{k\to\infty} P\{\xi_k^T \in A\} = \Pi^T(A) \qquad \forall\, A \in \mathcal{B} \,. \tag{3.3}$$

Let S be the set of global minima of $U(\cdot)$, i.e.,

$$S = \{x \in \Sigma : \; U(x) \leq U(y) \qquad \forall\, y \in \Sigma\}$$

(assume $S \neq \varnothing$ for the moment). Now for small T we expect $\Pi^T(\cdot)$ to be concentrated near S. Like the finite state case, the idea behind the general state annealing algorithm is that by choosing $T = T_k \to 0$ slowly enough the probability measure of ξ_k actually becomes concentrated near S.

Unlike the finite state case there are some technical problems in just verifying (3.3). We need to check Doeblin's condition and we also need a practical criterion to check whether $Q(\cdot,\cdot)$ is irreducible. These issues are investigated in 3.2. We will not use (3.3) in our analysis of the annealing algorithm with time-dependent temperature schedule. However (3.3) is of independent interest as it constitutes the theoretical justification of a continuous state version of the Metropolis algorithm which may be used for sampling from a continuous Gibbs distribution (c.f. [16]).

In 3.3, 3.4 we shall extend our result (Theorem 2.9) on the finite state annealing chain visiting S with probability one to the general state case, under essentially the condition that the state space Σ be a compact metric space and the energy function $U(\cdot)$ be continuous.

3.2. Ergodicity of the General State Annealing Chain at a Fixed Temperature

In this Section we shall discuss the ergodicity of the general state annealing chain at a fixed temperature. We shall use the notation of 3.1 except that we will fix a temperature schedule $T_k = T$, a positive constant, and suppress the dependence of the various quantities on T and also on the time index k whenever possible. In this notation we shall give conditions under which

$$\lim_{k\to\infty} P\{\xi_k \in A\} = \Pi(A) \qquad \forall\, A \in \mathcal{B} \,. \tag{3.4}$$

We have already remarked in 3.1 that (3.4) will hold if $Q(\cdot,\cdot)$ is irreducible and symmetric and a certain condition due to Doeblin is satisfied. Doeblin's condition will be satisfied if

(D) $0 < \phi(\Sigma) < \infty$ and there exists $\epsilon > 0$ such that $P(x,A) \leq 1 - \epsilon$ for all $x \in \Sigma$ and $A \in B$ with $\phi(A) < \epsilon$.

Under suitable conditions on Σ, $\phi(\cdot)$, $U(\cdot)$, and $q(\cdot,\cdot)$ we shall verify (D) and give a convenient characterization of the irreducibility of $Q(\cdot,\cdot)$. These same conditions will be used in 3.4 to analyze the general state annealing chain with time-dependent temperature schedule. We shall also give an example of a class of $q(\cdot,\cdot)$ which satisfy the stated conditions.

Consider the following set of conditions:

(A1) (Σ,ρ) is a compact metric space

(A2) (Σ,B,ϕ) is a nontrivial finite measure space with B the Borel subsets of Σ

(A3) $\phi(\cdot)$ is positive on open subsets of Σ

(A4) $U(\cdot)$ is continuous

(A5) $q(\cdot,\cdot)$ is bounded

(A6) $q(\cdot,\cdot)$ is continuous on $\{(x,y) \in \Sigma \times \Sigma : \ q(x,y) > 0\}$

(A7) $\phi(\{\cdot\})$ is lower semicontinuous on $\{x \in \Sigma : \ q(x,x) > 0\}$

We remark that not all of these conditions will be used to obtain every result.

The following proposition deals with Condition (D).

Proposition 3.1 Assume that (A1), (A2), (A4), (A5) hold. Then there exists $\epsilon > 0$ such that $P(x,A) \leq 1 - \epsilon$ for all $x \in \Sigma$ and $A \in B$ with $\phi(A) < \epsilon$.

Proof Using (3.2) and (A5) there exists a constant c_1 such that

$$P(x,A) \leq c_1 \phi(A) + \gamma(x) \qquad \forall\, x \in \Sigma\,, \ \forall\, A \in B\,. \tag{3.4}$$

Now

$$\begin{aligned}
\gamma(x) &= 1 - \int q(x,y)\, s(x,y)\, \phi(dy) \\
&\leq 1 - \int q(x,y) \exp\left[- \frac{|U(y) - U(x)|}{T} \right] \phi(dy) \\
&\leq 1 - c_2 \int q(x,y)\, \phi(dy) \\
&= 1 - c_2 \qquad\qquad \forall\, x \in \Sigma\,,
\end{aligned} \tag{3.5}$$

for some constant $c_2 > 0$, since (A1) and (A4) imply that $U(\cdot)$ is bounded. The Proposition now follows from (3.4) and (3.5). □

We next develop a criterion for the irreducibility of $Q(\cdot,\cdot)$ motivated by the finite state case. We shall say that given states x and y, x can *reach* y if there exists a sequence of states $x = x_0, \ldots, x_p = y$ such that $q(x_n, x_{n+1}) > 0$ for all $n = 0,\ldots,p-1$. Suppose that Σ is finite, $\phi(\cdot)$ is counting measure, and $q_{ij} = q(i,j)$. Then this definition reduces to that given in Chapter 2. Now the stochastic transition matrix $Q = [q_{ij}]$ is irreducible iff i can reach j for all $i,j \in \Sigma$. The following Theorem gives a similar criterion for the stochastic transition function $Q(x,A) = \int_A q(x,y)\, \phi(dy)$.

Theorem 3.1 Assume that (A1)-(A3), (A6) hold. Then $Q(\cdot,\cdot)$ is irreducible iff x can reach y for all $x \in \Sigma$ and ϕ-a.e. $y \in \Sigma$.

Proof Suppose that $Q(\cdot,\cdot)$ is irreducible and there exists $x \in \Sigma$ and $A \in \mathcal{B}$ with $\phi(A) > 0$ such that x cannot reach y for all $y \in A$. Then there exists a $d \in \mathbb{N}$ such that $Q^{(d)}(x,A) > 0$, and by Fubini's Theorem

$$\begin{aligned} Q^{(d)}(x,A) &= \int Q(x,dx_1) \cdots \int Q(x_{d-2},dx_{d-1})\, Q(x_{d-1},A) \\ &= \int q(x,x_1)\, \phi(dx_1) \cdots \int q(x_{d-2},x_{d-1})\, \phi(dx_{d-1}) \int_A q(x_{d-1},x_d)\, \phi(dx_d) \\ &= \int_{\Sigma^{d-1} \times A} q(x,x_1) \cdot \cdots \cdot q(x_{d-1},x_d)\, \phi^d(dx_1 \cdots dx_d) \\ &> 0 . \end{aligned} \tag{3.6}$$

Hence, $q(x,x_1),\ldots,q(x_{d-1},x_d) > 0$ for some $x_1,\ldots,x_{d-1} \in \Sigma$ and $x_d \in A$, and so x can reach some $y \in A$, a contradiction.

Conversely, suppose that x can reach y for all $x \in \Sigma$ and ϕ-a.e. $y \in \Sigma$. We first show that given $\epsilon > 0$ there exists a compact $C \subset \Sigma$ with $\phi(C) > \phi(\Sigma) - \epsilon$ such that x can reach y for all $x \in \Sigma$ and $y \in C$. Let $B \in \mathcal{B}$ such that $\phi(B) = \phi(\Sigma)$ and x can reach y for all $x \in \Sigma$ and $y \in B$. Recall that a Borel measure μ is regular if given $\delta > 0$ and a Borel set F there exists a compact set K and an open set G such that $K \subset F \subset G$ and $\mu(G) - \mu(K) < \delta$. It is known that finite Borel measures on compact metric spaces are regular (c.f. [28, Ch. 14] for a discussion of these matters). Hence by (A1), (A2) $\phi(\cdot)$ is a regular Borel measure and so there exists a compact $C \subset B$ such that $\phi(C) > \phi(B) - \epsilon > \phi(\Sigma) - \epsilon$ and necessarily x can reach y for all $x \in \Sigma$ and $y \in C$.

We next show that there exists a $d_1 \in \mathbb{N}$ such that x can reach y in not greater than d_1 steps for all $x \in \Sigma$ and $y \in C$. By (A6) if x can reach y in d(x,y) steps then there exists neighborhoods U_x of x and V_y of y such that u can reach v in d(x,y) steps for all $u \in U_x$ and $v \in V_y$. Now $\{U_x \times (V_y \cap C) : x \in \Sigma, y \in C\}$ is an open cover of compact $\Sigma \times C$ (in the relative topology) and so there exists $x_1,\ldots,x_N \in \Sigma$ and $y_1,\ldots,y_N \in C$ such that

$$\Sigma \times C \subset \bigcup_{n=1}^{N} U_{x_n} \times V_{y_n} .$$

Let

$$d_1 = \max_{n=1,\ldots,N} d(x_n,y_n) .$$

Now fix $x \in \Sigma$ and $A \in \mathcal{B}$ such that $\phi(A) > 0$. Ultimately we want to show that there exists $d \in \mathbb{N}$ such that $Q^{(d)}(x,A) > 0$. If $\phi(A) = \phi(\Sigma)$ then $Q^{(d)}(x,A) = 1$ for all $d \in \mathbb{N}$. So assume that $0 < \phi(A) < \phi(\Sigma)$. The next step is to show that there exists $d_2 \in \mathbb{N}$ and $D \in \mathcal{B}$ with $D \subset A$ and $\phi(D) > 0$ such that x can reach y in d_2 steps for all $y \in D$. Choose $0 < \epsilon < \phi(\Sigma) - \phi(A)$ in the definition of C above. Then $\phi(C \cap A) > \phi(\Sigma) - \phi(A) - \epsilon > 0$ and x can reach y in not greater than d_1 steps for all $y \in C \cap A$. For $n = 1,\ldots,d_1$ let

$$C_n = \{y \in C \cap A : \text{ x can reach y in n-steps}\}$$

Then $C_n \in \mathcal{B}$ for $n = 1,\ldots,d_1$ and

$$\bigcup_{n=1}^{d_1} C_n = C \cap A .$$

Hence since $\phi(C \cap A) > 0$ we may choose $d_2 \in \{1,\ldots,d_1\}$ such that $\phi(C_{d_2}) > 0$. Let $D = C_{d_2}$.

Let $d = d_2$. By one additional application of Fubini's Theorem to (3.6)

$$Q^{(d)}(x,A) \geq Q^{(d)}(x,D) = \int_D f(y)\, \phi(dy) \tag{3.7}$$

where

$$f(y) = \int q(x,x_1) \cdot \cdots \cdot q(x_{d-1},y)\, \phi^{d-1}(dx_1 \cdots dx_{d-1}) .$$

Since $f(\cdot)$ is a $\mathcal{B}$-measurable function on Σ and $\phi(D) > 0$, if we can show that $f(\cdot)$ is positive on D then by (3.7) $Q^{(d)}(x,A) > 0$ and we are done. We now show that $f(\cdot)$ is indeed positive on D. Fix $y \in D$ and let

$$\tilde{q}(x_1,\ldots,x_{d-1}) = q(x,x_1) \cdot \cdots \cdot q(x_{d-1},y) \qquad \forall\, x_1,\ldots,x_{d-1}\in\Sigma\,.$$

Then

$$f(y) = \int \tilde{q}(x_1,\ldots,x_{d-1})\, \phi^{d-1}(dx_1 \cdots dx_{d-1})\,.$$

Since x can reach y in d steps there exists $x_1,\ldots,x_{d-1}\in\Sigma$ such that $\tilde{q}(x_1,\ldots,x_{d-1}) > 0$. Using (A6) there exists neighborhoods B_n of x_n, $n = 1,\ldots,d-1$, such that $\tilde{q}(\cdot)$ is positive on $B_1\times \cdots \times B_{d-1}$. Since $\tilde{q}(\cdot)$ is a B^{d-1}-measurable function on Σ^{d+1} and $\phi^{d-1}(B_1\times \cdots \times B_{d-1}) = \phi(B_1) \cdot \cdots \cdot \phi(B_{d-1}) > 0$ by (A3), we have that $f(y) > 0$, and since $y\in D$ was chosen arbitrarily, we have $f(\cdot)$ is positive on D as required. □

We end this Section by giving an example of a class of $q(\cdot,\cdot)$ which have the property that the corresponding annealing chain makes "small" moves in a topological sense. This is consistent with the approach taken in the finite state case as discussed in Chapter 1. Of course if Σ is a metric space than the notion of smallness is well-defined. We construct a function $q(\cdot,\cdot)$ as follows. Assume that (A1)-(A3) hold, and let $p(\cdot,\cdot)$ and $R(\cdot)$ be positive continuous functions on $\Sigma\times\Sigma$ and Σ, respectively. Let

$$q(x,y) = c(x)\, p(x,y)\, \chi_{B(x,R(x))}(y) \qquad \forall\, x,y\in\Sigma\,, \tag{3.8}$$

where

$$c(x) = \left(\int_{B(x,R(x))} p(x,y)\, \phi(dy) \right)^{-1} \qquad \forall\, x\in\Sigma\,.$$

Note that if

$$\int p(x,y)\, \phi(dy) = 1 \qquad \forall\, x\in\Sigma\,,$$

and ξ is a random variable which density $p(x,\cdot)$ with respect to $\phi(\cdot)$, then $q(x,\cdot)$ is a density for the conditional distribution of ξ given $\xi\in B(x,R(x))$. The following proposition establishes that $q(\cdot,\cdot)$ satisfies (A5), (A6).

Proposition 3.2 Suppose that

$$\phi(\{y\in\Sigma :\ \rho(x,y) = R(x)\}) = 0 \qquad \forall\, x\in\Sigma\,. \tag{3.9}$$

Then $q(\cdot,\cdot)$ is bounded and continuous on $\{(x,y)\in\Sigma\times\Sigma :\ q(x,y) > 0\}$.

Proof Let

$$f(x,y) = p(x,y)\ \chi_{B(x,R(x))}(y) \qquad \forall\ x,y\in\Sigma\ ,$$

so that

$$q(x,y) = c(x)\ f(x,y) \qquad \forall\ x,y\in\Sigma\ ,$$

and

$$c(x) = \left(\int f(x,y)\ \phi(dy)\right)^{-1} \qquad \forall\ x\in\Sigma\ .$$

Using the continuity of $p(\cdot,\cdot)$ $R(\cdot)$, and $\rho(\cdot,\cdot)$ we have that $f(\cdot,\cdot)$ is a continuous function on $\{(x,y)\in\Sigma\times\Sigma:\ \rho(x,y) \neq R(x)\}$ and hence on $\{(x,y)\in\Sigma\times\Sigma:\ q(x,y) > 0\}$. We now show that $c(\cdot)$ is a continuous function on Σ. Let $x\in\Sigma$ and $\{x_n\}$ be a sequence in Σ such that $x_n\rightarrow x$. Then $f(x_n,y)\rightarrow f(x,y)$ for all $y\in\Sigma$ such that $\rho(x,y) \neq R(x)$. Hence by (3.9) $f(x_n,y)\rightarrow f(x,y)$ for ϕ-a.e. $y\in\Sigma$, and by the Dominated Convergence Theorem $c(x_n)\rightarrow c(x)$. Since x and $\{x_n\}$ were arbitrary, $c(\cdot)$ is continuous. The Proposition follows. □

Remark If Σ is a subset of $\mathbf{R}^r$, $\phi(\cdot)$ is Lebesgue measure and $\rho(x,y) = |y - x|$ then (3.9) is of course satisfied.

3.3 Asymptotic Analysis of a Class of Nonstationary Markov Chains

In this Section we analyze certain asymptotic properties of a class of nonstationary Markov chains. These chains have the property that their 1-step transition probabilities satisfy bounds similar to those satisfied by the d-step transition probabilities of the annealing chain. The results of this Section will be used in 3.4 to deduce corresponding asymptotic properties of the annealing chain.

We shall consider the following class of Markov chains. Let (Σ,ρ) be a compact metric space and (Σ,B,ϕ) a finite measure space with B the Borel subsets of Σ and $\phi(\cdot)$ positive on the open subsets of Σ. Let $\alpha(\cdot,\cdot)$ be a $[0,\infty]$-valued upper semicontinuous function on $\Sigma\times\Sigma$ and $\{\theta_k\}$ a sequence of real numbers with $0 < \theta_k \leq 1$. Let $\{\xi_k\}$ be a Markov chain with state space Σ and 1-step transition functions $\{P_k(\cdot,\cdot)\}$ whose ϕ-absolutely continuous components have densities $\{p_k(\cdot,\cdot)\}$ with the following property: for every $u,v\in\Sigma$ there exists a neighborhood $B_{u,v}$ of (u,v) in $\Sigma\times\Sigma$ and a positive number $K(u,v)$ such that

$$p_k(x,y) \geq K(u,v)\ \theta_k^{\alpha(u,v)} \qquad \forall\ (x,y)\in B_{u,v}\ . \tag{3.10}$$

Note that we do not assume there exist a positive number A such that

$$p_k(x,y) \geq A\ \theta_k^{\alpha(x,y)} \qquad \forall\ x,y\in\Sigma\ , \tag{3.11}$$

which is similar to (2.13). Of course if Σ is finite and $\phi(\cdot)$ is counting measure, then we do obtain (3.11).

The following theorem gives sufficient conditions under which $\{\xi_k\}$ visits an open subset of Σ infinitely often with probability one.

Theorem 3.2 Let Y be an open subset of Σ and

$$a = \sup_{x\in\Sigma\setminus Y}\ \inf_{y\in Y}\ \alpha(x,y) < \infty\ .$$

Suppose there exists $\epsilon > 0$ such that

$$\sum_{k=1}^{\infty}\ \theta_k^{a+\epsilon} = \infty\ . \tag{3.12}$$

Then $P\{\xi_k\in Y\ \text{i.o.}\} = 1$.

Remark If Σ is finite and $\phi(\cdot)$ is counting measure we obtain Theorem 2.4 modulo the factor of ϵ in (3.12) as compared with (2.40). However Theorem 3.2 cannot be proved by the simple argument used to prove Theorem 2.4, essentially because we assume only (3.10) and not (3.11).

We will need the following two lemmas for the proof of Theorem 3.2.

Lemma 3.1 Let $c > 0$. Then there exists a nonnegative lower semicontinuous function $L(\cdot,\cdot)$ on $\Sigma\times\Sigma$ with $L(x,y) > 0$ whenever $\alpha(x,y) < c$ and

$$p_k(x,y) \geq L(x,y)\ \theta_k^c \qquad \forall\ x,y\in\Sigma\ .$$

Proof Let $U = \{(u,v)\in\Sigma\times\Sigma :\ \alpha(u,v) < c\}$ which is an open subset of $\Sigma\times\Sigma$ since $\alpha(\cdot,\cdot)$ is upper semicontinuous on $\Sigma\times\Sigma$. Now by (3.10)

$$p_k(x,y) \geq K(u,v)\ \theta_k^c\ , \qquad \forall (x,y)\in B_{u,v},\ \forall (u,v)\in U\ .$$

Let

$$K_1(x,y) = \sup_{\substack{(u,v)\in U\ :\\ B_{u,v}\ni(x,y)}} K(u,v) \qquad \forall\ (x,y)\in U\ .$$

It follows that

$$p_k(x,y) \geq K_1(x,y)\ \theta_k^c \qquad \forall\ (x,y)\in U\ , \tag{3.13}$$

and

$$\inf_{(u,v)\in B_{x,y}} K_1(u,v) \geq K(x,y) > 0 \qquad \forall\ (x,y)\in U\ . \tag{3.14}$$

Let $K_2(\cdot,\cdot)$ be the *lower envelope* of $K_1(\cdot,\cdot)$, i.e.,

$$K_2(x,y) = \sup_{\delta>0} \inf_{0<\rho(u,x)+\rho(v,y)<\delta} K_1(u,v) \qquad \forall\ x,y\in U\ .$$

Then (c.f. [1]) $K_2(\cdot,\cdot)$ is a lower semicontinuous function on U and $K_2(x,y) \leq K_1(x,y)$ for all $(x,y)\in U$. Also (3.14) implies that $K_2(\cdot,\cdot)$ is positive. Let

$$L(x,y) = \begin{cases} K_2(x,y) & \text{if } (x,y)\in U \\ 0 & \text{if } (x,y)\notin U\ , \end{cases}$$

for all $x,y\in\Sigma$. Since U is open and $K_2(\cdot,\cdot)$ is a positive lower semicontinuous function on U, $L(\cdot,\cdot)$ is a lower semicontinuous function on $\Sigma\times\Sigma$ which is positive on U. Furthermore

$$K_1(x,y) \geq K_2(x,y) = L(x,y) \qquad \forall\ (x,y)\in U\ . \tag{3.15}$$

The Lemma now follows from (3.13) and (3.15). □

Lemma 3.2 Let Y be an open subset of Σ and

$$a = \sup_{x\in\Sigma\setminus Y} \inf_{y\in Y} \alpha(x,y)\ .$$

Let $\epsilon > 0$ and $L(\cdot,\cdot)$ be a lower semicontinuous function on $\Sigma\times\Sigma$ such that $L(x,y) > 0$ whenever $\alpha(x,y) < a + \epsilon$. Then there exists open sets $W_1,\ldots,W_M$ contained in Y such that

$$\sup_{x\in\Sigma\setminus Y} \min_{m=1,\ldots,M} \sup_{y\in W_m} \alpha(x,y) \leq a + \epsilon\ ,$$

$$\inf_{x\in\Sigma\setminus Y} \max_{m=1,\ldots,M} \inf_{y\in W_m} L(x,y) > 0\ .$$

Proof Let $X = \Sigma\setminus Y$. We first show there exists a relatively open cover $U_1,\ldots,U_N$ of X and open sets $V_1,\ldots,V_N$ contained in Y such that $\alpha(x,y) < a + \epsilon$ for all $x\in U_n$, $y\in V_n$, and $n = 1,\ldots,N$. For every $x\in X$ there exists a $y\in Y$ such that $\alpha(x,y) < a + \epsilon$, and since $\alpha(\cdot,\cdot)$ is upper semicontinuous there exists neighborhoods A_x of x and B_x of y such that $\alpha(u,v) < a + \epsilon$ for all $u\in A_x$ and $v\in B_x$. Since $\{A_x\cap X : x\in X\}$ is an open cover of compact X (in the relative topology), there exists $x_1,\ldots,x_N\in X$ such that

$$X \subset \bigcup_{n=1}^{N} A_{x_n} .$$

Let $U_n = A_{x_n} \cap X$ and $V_n = B_{x_n} \cap Y$ for $n = 1,...,N$.

We next show there exists a $\delta > 0$ such that for every $x \in X$ there exists a $y \in Y$ and an $n \in \{1,...,N\}$ such that $x \in U_n$, $y \in V_n$, and $L(x,y) > \delta$. Let

$$f_n(x) = \sup_{y \in V_n} L(x,y) \qquad \forall\, x \in X\,,\ \forall\, n = 1,...,N$$

and

$$f(x) = \max_{\substack{n=1,...,N: \\ U_n \ni x}} f_n(x) \qquad \forall\, x \in X\,.$$

Since $L(\cdot,\cdot)$ is lower semicontinuous, $\{f_1(\cdot),...,f_N(\cdot)\}$ are lower semicontinuous functions on X, and since $\{U_1,...,U_N\}$ are open in X, $f(\cdot)$ is a lower semicontinuous function on X. Now $L(x,y) > 0$ whenever $\alpha(x,y) < a + \epsilon$ and in particular when $x \in U_n$ and $y \in V_n$ for some $n = 1,...,N$. It follows that $f(\cdot)$ is positive. Hence since $f(\cdot)$ is a positive lower semicontinuous function on compact X we can choose

$$0 < \delta < \inf_{x \in X} f(x)\,.$$

Combining the above results, for every $x \in X$ there exists a $y \in Y$ such that $\alpha(x,y) < a + \epsilon$ and $L(x,y) > \delta$. We can now find similarly to the construction of $U_1,...,U_N$ and $V_1,...,V_N$ above a cover $\tilde{U}_1,...,\tilde{U}_M$ of X and open $\tilde{V}_1,...,\tilde{V}_M$ contained in Y such that $\alpha(x,y) < a + \epsilon$ and $L(x,y) > \delta$ for all $x \in \tilde{U}_m$, $y \in \tilde{V}_m$, and $m = 1,...,M$. Let $W_m = \tilde{V}_m$ for $m = 1,...,M$ to complete the proof of the Lemma. □

Proof of Theorem 3.2 Let $X = \Sigma \backslash Y$. From Lemmas 3.1 and 3.2 there exists a $\delta > 0$ and open sets $W_1,...,W_M$ contained in Y such that for every $x \in X$ there exists an $m \in \{1,...,M\}$ such that

$$p_k(x,y) \geq \delta\, \theta_k^{a+\epsilon} \qquad \forall\, y \in W_m\,. \tag{3.16}$$

Using (3.16) and the Markov property

$$P \bigcap_{k=m}^{n} \{\xi_k \in X\} \leq P\{\xi_m \in X\} \prod_{k=m}^{n-1} \sup_{x \in X} P\{\xi_{k+1} \in X | \xi_k = x\}$$

$$\leq \prod_{k=m}^{n-1} \left[1 - \inf_{x \in X} P\{\xi_{k+1} \in Y | \xi_k = x\}\right]$$

$$\leq \prod_{k=m}^{n-1} \left[1 - \inf_{x \in X} \int_Y p_k(x,y)\, \phi(dy)\right]$$

$$\leq \prod_{k=m}^{n-1} \left[1 - A\, \theta_k^{a+\epsilon}\right] \qquad \forall\, n > m\,,$$

where $A = \delta \cdot \min_{m=1,\ldots,M} \phi(W_m) > 0$ (since $\phi(\cdot)$ is assumed positive on open subsets of Σ). Hence

$$P \bigcap_{k=m}^{\infty} \{\xi_k \in X\} \leq \prod_{k=m}^{\infty} \left[1 - A\, \theta_k^{a+\epsilon}\right] = 0 \qquad \forall\, m\,,$$

where the divergence of the infinite product follows from the divergence of the infinite sum (3.12), and the Theorem follows. □

3.4 Convergence of the General State Annealing Algorithm

In the Section we apply the results of 3.3 to obtain certain asymptotic properties of the general state annealing algorithm. Throughout this Section (3.4) we shall use the notation introduced in 3.1. We shall also refer to conditions (A1)-(A7) given in 3.2.

3.4.1 Bounds on Transition Probabilities for the General State Annealing Chain

In order to apply the results of 3.3 we need to obtain a bound on the d-step transition density $p^{(k,k+d)}(\cdot,\cdot)$ of the ϕ-absolutely continuous component of the d-step transition function $P^{(k,k+d)}(\cdot,\cdot)$ of the annealing chain $\{\xi_k\}$. Toward this end we make the following definitions. For every $x,y \in \Sigma$ and $d \in \mathbb{N}$ let $\Lambda_d(x,y)$ be the subset of Σ^{d+1} such that $(x = x_0,\ldots,x_d = y) \in \Lambda_d(x,y)$ if for every $n = 0,\ldots,d-1$ one of the following is true:

(i) $x_{n+1} \neq x_n$ and $q(x_n,x_{n+1}) > 0$

(ii) $x_{n+1} = x_n$ and $q(x_n,x_{n+1}) > 0$, $\phi(\{x_n\}) > 0$

(iii) $x_{n+1} = x_n$ and $q(x_n,z) > 0$ for some $z \in \Sigma$ with $U(z) > U(x_n)$.

The following proposition gives an alternative characterization of $\Lambda_d(\cdot,\cdot)$.

Proposition 3.3 Assume (A1)-(A6). Let $x,y \in \Sigma$ and $d \in \mathbb{N}$ Then $(x = x_0,\ldots,x_d = y) \in \Lambda_d(x,y)$ iff there exists a version of $p_k(\cdot,\cdot)$ such that

$$\max \{p_k(x_n,x_{n+1}),\ P_k(x_n,\{x_{n+1}\})\} > 0 \qquad \forall\ n = 0,\ldots,d-1\ .$$

Proof By the Radon-Nikodym Theorem and (3.2)

$$\begin{aligned} P_k(x,A) &= \int_A p_k(x,y)\ \phi(dy) + \tilde{P}_k(x,A) \\ &= \int_A q(x,y)\ s_k(x,y)\ \phi(dy) + \gamma_k(x)\ \delta(x,A) \end{aligned}$$

for all $x \in \Sigma$ and $A \in B$, where $\phi(\cdot)$ and $\tilde{P}_k(x,\cdot)$ are mutually singular. Hence

$$p_k(x,y) = q(x,y)\ s_k(x,y) + \frac{\gamma_k(x)}{\phi(\{x\})}\ \chi_{\{x\}}(y)$$

for all $x \in \Sigma$ and ϕ-a.e. $y \in \Sigma$, and

$$P_k(x,\{y\}) = q(x,y)\ s_k(x,y)\ \phi(\{y\}) + \gamma_k(x)\ \chi_{\{x\}}(y) \tag{3.17}$$

for all $x,y \in \Sigma$. Fix the following version of $p_k(\cdot,\cdot)$:

$$\bar{p}_k(x,y) = \begin{cases} q(x,y)\ s_k(x,y) & \text{if } y \neq x\ , \\ q(x,x) + \dfrac{\gamma_k(x)}{\phi(\{x\})} & \text{if } y = x,\ \phi(\{x\}) > 0 \\ 0 & \text{if } y = x\ ,\ \phi(\{x\}) = 0 \end{cases} \tag{3.18}$$

for all $x,y \in \Sigma$. Now under (A1)-(A6) for every $x \in \Sigma$ $\gamma_k(x) > 0$ iff $q(x,z) > 0$ for some $z \in \Sigma$ with $U(z) > U(x)$. It follows from (3.17), (3.18) and this last remark that (i)-(iii) hold iff $\bar{p}_k(x_n,x_{n+1}) > 0$ or $P_k(x_n,\{x_{n+1}\}) > 0$ for all $n = 0,\ldots,d-1$. □

Suppose Σ is finite, $\phi(\cdot)$ is counting measure and $q_{ij} = q(i,j)$. In view of Proposition 3.3 the above definition of $\Lambda_d(\cdot,\cdot)$ reduces to that given in Chapter 2.

For each $d \in \mathbb{N}$ let

$$U_d(x_0,\ldots,x_d) = \sum_{n=0}^{d-1} \max\{0,\, U(x_{n+1}) - U(x_n)\}\ ,$$

for all $x_0,\ldots,x_d \in \Sigma$, and

$$V_d(x,y) = \begin{cases} \inf\limits_{\lambda \in \Lambda_d(x,y)} U_d(\lambda) & \text{if } y \neq x \\ \infty & \text{if } y = x\ , \end{cases} \tag{3.19}$$

$$V(x,y) = \inf_d V_d(x,y)\ ,$$

for all $x,y \in \Sigma$. We shall call $V_d(x,y)$ the d-*step transition energy from* x *to* y, and $V(x,y)$ the *transition energy from* x *to* y. We should like to point out a difference in the definition of $V_d(\cdot,\cdot)$ here and in Chapter 2 (compare (3.19) and (2.45)). Here we set $V_d(x,x) = \infty$; see the remark following Proposition 3.4 for an explanation.

We first prove that the d-step transition energy (and hence the transition energy itself) is an upper semicontinuous function.

Proposition 3.4 Assume (A1)-(A6). Then $V_d(\cdot,\cdot)$ is an upper semicontinuous function from $\Sigma\times\Sigma$ into $[0,\infty]$.

Proof Let $x,y \in \Sigma$ such that $V_d(x,y) < \infty$, and let $\epsilon > 0$. From (3.19) we have that $y \neq x$ and there exists $\lambda \in \Lambda_d(x,y)$ such that $U_d(\lambda) < V_d(x,y) + \epsilon/2$. It is clear that λ can be chosen such that all of the self-transitions in λ occur consecutively. We consider here the following case (the other cases are similar):

$$\lambda = (x = x_0 = \cdots = x_{m-1} \neq x_m \neq \cdots \neq x_d = y)$$

where $1 < m < d$. Now $q(x,x) > 0$ or $q(x,z) > 0$ for some $z \in \Sigma$ with $U(z) > U(x)$, and $q(x_n,x_{n+1}) > 0$ for all $n = m-1,\ldots,d-1$. Hence by (A4), (A6) we can choose neighborhoods B_x of x and B_y of y with $B_x \cap B_y = \varnothing$ such that for every $u \in B_x$ and $v \in B_y$ we have $q(u,u) > 0$ or $q(u,z) > 0$, $q(u,x_m) > 0$, $q(x_{d-1},v) > 0$, and

$$|U(u) - U(x)| + |U(v) - U(y)| < \frac{\epsilon}{2}\ .$$

Now let $(u,v) \in B_x \times B_y$ and

$$\sigma = (u,\ldots,u,x_m,\ldots,x_{d-1},v)\ .$$

m times

Then $\sigma\in\Lambda_d(u,v)$ and

$$|U_d(\sigma) - U_d(\lambda)| \le |U(u) - U(x)| + |U(v) - U(y)| < \frac{\epsilon}{2}$$

and so

$$V_d(u,v) \le U_d(\sigma) \le U_d(\lambda) + \frac{\epsilon}{2} < V_d(x,y) + \epsilon$$

and consequently $V_d(\cdot,\cdot)$ is upper semicontinuous at (x,y). Since x,y were arbitrary points in Σ which satisfy $V_d(x,y) < \infty$, $V_d(\cdot,\cdot)$ is upper semicontinuous. □

Remark Let

$$\tilde{V}_d(x,y) = \inf_{\lambda\in\Lambda_d(x,y)} U_d(\lambda) \qquad \forall\ x,y\in\Sigma$$

so that $\tilde{V}_d(x,y) = V_d(x,y)$ for $y \neq x$ but $\tilde{V}_d(x,x) \neq V_d(x,x)$ in general. It is easy to construct examples such that $\tilde{V}_d(x,y)$ is not upper semicontinuous at $y = x$. We defined $V_d(x,x) = \infty$ to avoid this problem.

The following theorem gives a lower bound on the d-step transition probabilities of the annealing chain in terms of the d-step transition energy.

Theorem 3.3 Assume (A1)-(A7). Let $\{T_k\}$ be monotone nonincreasing and $d\in\mathbb{N}$ Then there exists a version of $p^{(k,k+d)}(\cdot,\cdot)$ with the following property: given $\epsilon > 0$ for every $u,v\in\Sigma$ there exists a neighborhood $B_{u,v}$ of (u,v) in $\Sigma\times\Sigma$ and a positive number $K(u,v)$ such that

$$p^{(k,k+d)}(x,y) \ge K(u,v)\ \exp\left[-\frac{V_d(u,v) + \epsilon}{T_{k+d-1}}\right] \qquad \forall\ (x,y)\in B_{u,v}\ . \tag{3.20}$$

Remark We do not assert (nor do we believe it is true in general) that there exists a positive number A such that

$$p^{(k,k+d)}(x,y) \ge A\ \exp\left[-\frac{V_d(x,y) + \epsilon}{T_{k+d-1}}\right] \qquad \forall x,y\in\Sigma\ , \tag{3.21}$$

which is similar to the lower bound in (2.48) for the finite state case. Of course if Σ is finite and $\phi(\cdot)$ is counting measure than we do obtain (3.21).

We will need the following lemma for the proof of Theorem 3.3.

Lemma 3.3 Assume (A1)-(A7). Then $P_k(\cdot,\{\cdot\})$ is a continuous function on Σ.

Proof From (3.2)

$$\begin{aligned} P_k(x,\{x\}) &= q(x,x)\ \phi(\{x\}) + \gamma_k(x) \\ &= 1 - \int_{y\neq x} q(x,y)\ s_k(x,y)\ \phi(dy) \\ &= q(x,x)\ \phi(\{x\}) + \int q(x,y)\ [1 - s_k(x,y)]\ \phi(dy)\ , \end{aligned} \tag{3.22}$$

for all $x\in\Sigma$. Let $x\in\Sigma$ and $\{x_n\}$ be a sequence in Σ with $x_n\longrightarrow x$. Now from the second equality in (3.22)

$$\begin{aligned} \limsup_{n\to\infty} P_k(x_n,\{x_n\}) &\leq 1 - \lim_{n\to\infty} \int_{\substack{y\neq x,\\ q(x,y)>0}} q(x_n,y)\ s_k(x_n,y)\ \phi(dy) \\ &= 1 - \int_{\substack{y\neq x,\\ q(x,y)>0}} q(x,y)\ s_k(x,y)\ \phi(dy) \\ &= P_k(x,\{x\})\ , \end{aligned} \tag{3.23}$$

where we have used (A1)-(A6) and the Dominated Convergence Theorem to evaluate the limit. Also, from the third equality in (3.22)

$$\begin{aligned} \liminf_{n\to\infty} P_k(x_n,\{x_n\}) &\geq \liminf_{n\to\infty} q(x_n,x_n)\ \phi(\{x_n\}) \\ &\quad + \lim_{n\to\infty} \int_{q(x,y)>0} q(x_n,y)\ [1 - s_k(x_n,y)]\ \phi(dy) \\ &\geq q(x,x)\ \phi(\{x\}) + \lim_{n\to\infty} \int_{q(x,y)>0} q(x_n,y)\ [1 - s_k(x_n,y)]\ \phi(dy) \\ &= q(x,x)\ \phi(\{x\}) + \int_{q(x,y)>0} q(x,y)\ [1 - s_k(x,y)]\ \phi(dy) \\ &= P_k(x,\{x\})\ , \end{aligned} \tag{3.24}$$

where we have used (A6), (A7) to obtain the second inequality and (A1)-(A6) and the Dominated Convergence Theorem to evaluate the limit. Combining (3.23) and (3.24) gives

$$\lim_{n\to\infty} P_k(x_n,\{x_n\}) = P_k(x,\{x\}) ,$$

and since x, $\{x_n\}$ were arbitrary the Lemma follows. □

Proof of Theorem 3.3

Let

$$r_k(x,y) = \begin{cases} q(x,y) & \text{if } y \neq x , \\ P_k(x,\{x\}) & \text{if } y = x , \end{cases} \tag{3.25}$$

for all $x,y\in\Sigma$, and

$$\tilde{r}_k(x_0,\ldots,x_d) = \prod_{n=0}^{d-1} r_{k+n}(x_n,x_{n+1}) , \tag{3.26}$$

$$\tilde{r}(x_0,\ldots,x_d) = \inf_k \tilde{r}_k(x_0,\ldots,x_d) , \tag{3.27}$$

for all $x_0,\ldots,x_d\in\Sigma$. If $\lambda\in\Sigma^{d+1}$ then since $\{T_k\}$ is nonincreasing $\{\tilde{r}_k(\lambda)\}$ is nondecreasing and so $\tilde{r}(\lambda) = \tilde{r}_1(\lambda)$ obtains the infimum. Note that $\tilde{r}(\lambda) > 0$ for all $\lambda\in\Lambda_d(x,y)$, $x,y\in\Sigma$.

For every $x\in\Sigma$ define a measure $\psi(x,\cdot)$ on (Σ,B) by

$$\psi(x,A) = \phi(A) + [1 - \phi(\{x\})]\ \delta(x,A)$$

and define a measure $\tilde{\psi}(x,\cdot)$ on (Σ^{d+1},B^{d+1}) by

$$\tilde{\psi}(x,A_0\times\cdots\times A_d) = \int_{A_0} \delta(x,dx_0) \int_{A_1} \psi(x_0,dx_1) \cdots \int_{A_d} \psi(x_{d-1},dx_d) .$$

It follows from (3.2) and (3.25)-(3.27) that

$$P^{(k,k+d)}(x,A) \geq \int_{\substack{\lambda\in\Lambda_d(x,y) , \\ y\in A}} \tilde{r}(\lambda) \exp\left[-\frac{U_d(\lambda)}{T_{k+d-1}}\right] \tilde{\psi}(x,d\lambda) \tag{3.28}$$

for all $x\in\Sigma$ and $A\in B$.

For every $n = 1,\ldots,d$ and $x,y\in\Sigma$ let

$$M_n(x,y) = \{(x_0,\ldots,x_d)\in\Lambda_d(x,y) :\ x_{n-1} \neq y\ ,\ x_k = y \qquad \forall\ k = n,\ldots,d\}$$

Then from (3.28)

$$P^{(k,k+d)}(x,A) \geq \sum_{n=1}^{d} \int_{\substack{\lambda\in M_n(x,y), \\ y\in A}} \tilde{r}(\lambda) \exp\left[-\frac{U_d(\lambda)}{T_{k+d-1}}\right] \tilde{\psi}(x,d\lambda) . \tag{3.29}$$

For every $n = 1,\ldots,d$ let $\Pi_n(\cdot)$ be the projection map from Σ^{d+1} to Σ^n, and for

every $x\in\Sigma$ let $\psi_n(x,\cdot)$ be the image measure of $\tilde{\psi}(x,\cdot)$ under $\Pi_n(\cdot)$; also let

$$x\cdot 1_n = (x,\ldots,x)\,.$$

n copies

Then applying Fubini's Theorem to (3.29)

$$P^{(k,k+d)}(x,A) \geq \int_A f_k(x,y)\ \phi(dy) \tag{3.30}$$

where

$$f_k(x,y) = \sum_{n=1}^{d-1} \int_{\Pi_n M_n(x,y)} \tilde{r}(\lambda, y\cdot 1_{d-n+1}) \exp\left[-\frac{U_n(\lambda)}{T_{k+d-1}}\right] \psi_n(x,d\lambda)\,. \tag{3.31}$$

Now by the Radon-Nikodym Theorem we have

$$P^{(k,k+d)}(x,A) = \int_A p^{(k,k+d)}(x,y)\ \phi(dy) + \tilde{P}^{(k,k+d)}(x,A) \tag{3.32}$$

where $\phi(\cdot)$ and $\tilde{P}^{(k,k+d)}(x,\cdot)$ are mutually singular. It follows from (3.30) and (3.32) that

$$\int_A p^{(k,k+d)}(x,y)\ \phi(dy) \geq \int_A f_k(x,y)\ \phi(dy)$$

for all $x\in\Sigma$ and $A\in B$, and so

$$p^{(k,k+d)}(x,y) \geq f_k(x,y) \tag{3.33}$$

for all $x\in\Sigma$ and ϕ-a.e. $y\in\Sigma$, and consequently there is a version of $p^{(k,k+d)}(\cdot,\cdot)$ such that (3.33) holds for all $x,y\in\Sigma$.

Fix $\epsilon > 0$ and $u,v\in\Sigma$. For each $x,y\in\Sigma$ if $V_d(u,v) < \infty$ let

$$N_n(x,y) = \{\lambda\in\Pi_n M_n(x,y):\ U_n(\lambda) < V_d(u,v) + \epsilon\}$$

for $n = 1,\ldots,d$, and set

$$g(x,y) = \sum_{n=1}^{d} \int_{N_n(x,y)} \tilde{r}(\lambda, y\cdot 1_{d-n+1})\ \psi_n(x,d\lambda)\ ; \tag{3.34}$$

if $V_d(u,v) = \infty$ set $g(x,y) = 1$. Then from (3.31)

$$f_k(x,y) \geq g(x,y) \exp\left[-\frac{V_d(u,v)+\epsilon}{T_{k+d-1}}\right] \qquad \forall\ x,y\in\Sigma\,. \tag{3.35}$$

We make the following

Claim There exists a neighborhood B of (u,v) in $\Sigma\times\Sigma$ such that

$$\inf_{(x,y)\in B} g(x,y) > 0 .$$

Suppose the Claim is true. Then by setting $B_{u,v} = B$ and

$$K(u,v) = \inf_{(x,y)\in B} g(x,y)$$

and combining (3.33) and (3.35) we obtain the Theorem. It remains to prove the Claim.

Proof of Claim Assume $V_d(u,v) < \infty$. From (3.19) there exists $\lambda\in\Lambda_d(u,v)$ such that $U_d(\lambda) < V_d(u,v) + \epsilon$, and since $\tilde{r}(\lambda) > 0$ there exists $\delta > 0$ such that $\tilde{r}(\lambda) > \delta$. Also from (3.19) we must have $u \neq v$, which implies there exists an $n\in\{1,...,d\}$ such that $\lambda\in M_n(u,v)$. It is clear λ can be chosen such that all of the self-transitions in λ which occur before the n^{th} transition (which is not a self-transition) occur consecutively. We consider here the following case (the other cases are similar):

$$\lambda = (u = u_0 = \cdots = u_{m-1} \neq u_m \neq \cdots \neq u_{n-1} \neq u_n = u_{n+1} = \cdots = v)$$

where $1 < m < n < d$. Using (A4), (A6) and Lemma 3.3 we can choose neighborhoods B_u of u, B_v of v, and $\bar{B}_k$ of u_k for $k = m,...,n-1$ with $\bar{B}_{n-1} \cap B_v = \emptyset$, such that for every $x\in B_u$ and $y\in B_v$ we have $U_d(\sigma) < V_d(u,v) + \epsilon$ and $\tilde{r}(\sigma) > \delta$ for all $\sigma\in\{x\}^m \times \bar{B}_m \times \cdots \times \bar{B}_{n-1} \times \{y\}^{d-n+1}$. Let

$$O_{x,y} = \{x\}^m \times \bar{B}_m \times \cdots \times \bar{B}_{n-1} \qquad \forall\, x,y\in\Sigma ,$$

and $B = B_u \times B_v$. Then for every $(x,y)\in B$ we have $O_{x,y}\subset N_n(x,y)$ and hence from (3.34)

$$\begin{aligned} g(x,y) &\geq \int_{O_{x,y}} \delta\, \psi_n(x,d\lambda) \\ &\geq \delta\, \phi(\bar{B}_m) \cdot \cdots \cdot \phi(\bar{B}_n) \\ &> 0 \end{aligned}$$

by (A3). This proves the Claim and hence the Theorem. □

3.4.2 Visiting of Neighborhood of the Set of Global Minima with Probability One

We now give a theorem whichs gives conditions such that annealing chain $\{\xi_k\}$ visits a given neighborhood of S infinitely often with probability one. Let $\epsilon > 0$ and

$$S_\epsilon = \{x \in \Sigma : \; U(x) < \inf_{y \in \Sigma} U(y) + \epsilon\}$$

To avoid trivialities we will need the following assumption:

(P) Every $i \in \Sigma \backslash S_\epsilon$ can reach some $j \in S_\epsilon$.

Let

$$V_\epsilon^* = \sup_{x \in \Sigma \backslash S_\epsilon} \inf_{y \in S_\epsilon} V(x,y) \; .$$

Note that under (A1)-(A4) and (A6) (so that $\Sigma \backslash S_\epsilon$ is compact and by Proposition 3.4 $V(\cdot,\cdot)$ is upper semicontinuous) (P) holds iff $V_\epsilon^* < \infty$.

Theorem 3.4 Assume (A1)-(A7) and (P). Let $\{T_k\}$ be monotone nonincreasing and

$$\sum_{k=1}^{\infty} \exp\left[-\frac{V_\epsilon^* + \delta}{T_k}\right] = \infty \tag{3.36}$$

for some $\delta > 0$. Then $P\{\xi_k \in S_\epsilon \text{ i.o.}\} = 1$ for all $\epsilon > 0$.

Proof We first show that there exists $d \in \mathbb{N}$ such that

$$V_\epsilon^* \geq \sup_{x \in \Sigma \backslash S_\epsilon} \inf_{y \in S_\epsilon} V_d(x,y) - \frac{\delta}{2} \tag{3.37}$$

For every $x \in \Sigma \backslash S_\epsilon$ there exists a $d(x) \in \mathbb{N}$ such that

$$\inf_{y \in S_\epsilon} V_{d(x)}(x,y) < \inf_{y \in S_\epsilon} V(x,y) + \frac{\delta}{2} \leq V_\epsilon^* + \frac{\delta}{2} \; .$$

But by Proposition 3.4 for every $x \in \Sigma \backslash S_\epsilon$ $V_{d(x)}(\cdot,\cdot)$ is an upper semicontinuous function on $\Sigma \times \Sigma$ and so $\inf_{y \in S_\epsilon} V_{d(x)}(\cdot,y)$ is an upper semicontinuous function on Σ, and consequently there exists a neighborhood B_x of x such that

$$\inf_{y \in S_\epsilon} V_{d(x)}(u,y) \leq V_\epsilon^* + \frac{\delta}{2} \qquad \forall \; u \in B_x \; .$$

Now $\{B_x \cap (\Sigma \backslash S_\epsilon) : \; x \in \Sigma \backslash S_\epsilon\}$ is an open cover of compact $\Sigma \backslash S_\epsilon$ (in the relative topology) and so there exists $x_1, \ldots, x_N \in \Sigma \backslash S_\epsilon$ such that

$$\Sigma \backslash S_\epsilon \subset \bigcup_{n=1}^{N} B_{x_n} .$$

Let $d^* = \max_{n=1,\ldots,N} d(x_n)$. Now it is easy to see that for every $x \in \Sigma$

$$\inf_{y \in S_\epsilon} V_n(x,y) \leq \inf_{y \in S_\epsilon} V_m(x,y) \qquad \forall\, n \geq m .$$

Hence for every $x \in \Sigma \backslash S_\epsilon$

$$\inf_{y \in S_\epsilon} V_{d^*}(x,y) = \min_{n \leq d^*} \inf_{y \in S_\epsilon} V_n(x,y) < V_\epsilon^* + \frac{\delta}{2}$$

and (3.37) follows by setting $d = d^*$.

Next, from Theorem 3.3 for every $u,v \in \Sigma$ there exists a neighborhood $B_{u,v}$ of (u,v) in $\Sigma \times \Sigma$ and a positive number $K(u,v) > 0$ such that

$$p^{(k,k+d)}(x,y) \geq K(u,v) \exp\left[- \frac{V_d(u,v) + \delta/2}{T_{k+d-1}} \right] \qquad \forall\, (x,y) \in B_{u,v} .$$

Let

$$\tilde{\xi}_k = \xi_{kd} , \quad \theta_k = \exp\left(- \frac{1}{T_{kd+d-1}} \right)$$

and

$$\alpha(x,y) = V_d(x,y) + \frac{\delta}{2} \qquad \forall\, x,y \in \Sigma . \tag{3.38}$$

Then $\{\tilde{\xi}_k\}$ is a Markov chain with 1-step transition functions $\{\tilde{P}_k(\cdot,\cdot)\}$ whose ϕ-absolutely continuous components have densities $\{\tilde{p}_k(\cdot,\cdot)\}$ which satisfy

$$\tilde{p}_k(x,y) \geq K(u,v)\, \theta_k^{\alpha(u,v)} \qquad \forall\, (x,y) \in B_{u,v},\ \forall\, u,v \in \Sigma .$$

Let

$$a = \sup_{x \in \Sigma \backslash S_\epsilon} \inf_{y \in S_\epsilon} \alpha(x,y) .$$

By (3.37) and (3.38) $a \leq V_\epsilon^* + \delta$. Hence since $\{T_k\}$ is nonincreasing the divergence of the series in (3.36) implies that

$$\sum_{k=1}^{\infty} \theta_k^a = \infty .$$

Hence we may apply Theorem 3.1 to $\{\tilde{\xi}_k\}$ with $Y = S_\epsilon$ to get $P\{\tilde{\xi}_k \in S_\epsilon \text{ i.o.}\} = 1$ and so $P\{\xi_k \in S_\epsilon \text{ i.o.}\} = 1$. □

Remark If Σ is finite, $\phi(\cdot)$ is counting measure, and ϵ is small enough we obtain Theorem 2.9 modulo the factor of δ in (3.36) as compared with (2.67).

CHAPTER IV
DIFFUSION TYPE ALGORITHMS

4.1 Introduction to the Langevin Algorithm

In Chapter 2 we discussed the annealing algorithm proposed by Kirkpatrick et. al. [19] and Cerny [3] for combinatorial optimization. In Chapter 3 we extended the annealing for optimization on general spaces. Motivated by image processing problems with continuous variables, Geman and independently Grenander [13] have recently proposed using diffusions for optimization on multidimensional Euclidean space. In this Section we describe this method. Like the annealing algorithm, this approach to global optimization has generated alot of interest and there already exists a significant literature on the subject.

Let $U(\cdot)$ be a nonnegative continuously differentiable function on $\mathbb{R}^r$. The goal is to find a point in $\mathbb{R}^r$ which *minimizes or nearly minimizes* $U(\cdot)$. Let $T(\cdot)$ be a positive Borel function on $[0,\infty)$. As with the annealing algorithm we shall refer to $U(\cdot)$ as the *energy function* and $T(\cdot)$ as the *temperature schedule*. Let $w(\cdot)$ be a standard r-dimensional Wiener process and let $x(\cdot)$ be a solution of the stochastic differential equation

$$dx(t) = -\nabla U(x(t))dt + \sqrt{2T(t)}\, dw(t)\,, \qquad t \geq 0\,, \tag{4.1}$$

for some initial condition $x(0) = x_0$ (by a solution we mean that $x(\cdot)$ is a separable process with continuous sample paths with probability one, $x(\cdot)$ is nonanticipative with respect to $w(\cdot)$, and $x(\cdot)$ satisfies the Ito integral equation corresponding to (4.1)). For a fixed temperature $T(t) = T > 0$, (4.1) is the Langevin equation, proposed by Langevin in 1908 to describe the motion of a particle in a viscous fluid. Geman and Grenander suggested that (4.1) could be used to minimize $U(\cdot)$ by letting $T(t) \to 0$. Following Gidas' [11] notation, we shall call the algorithm which simulates the sample paths of $x(\cdot)$ with $T(t) \to 0$ the *Langevin algorithm.*

The motivation behind the Langevin algorithm is similar to that of the annealing algorithm. Let $x^T(\cdot)$ be the solution of (4.1) with $T(t) = T$, a positive constant, and let $P^T(\cdot,\cdot,\cdot)$ be its (stationary) transition function, i.e.,

- for every $t \geq 0$ and $A \in \mathbb{B}^r$ $P^T(t,\cdot,A)$ is a Borel function on $\mathbb{R}^r$
- for every $t \geq 0$ and $x \in \mathbb{R}^r$ $P^T(t,x,\cdot)$ is a probability measure on $(\mathbb{R}^r,\mathbb{B}^r)$
- $P^T(t,x,A) = \int P^T(s,x,dy)\, P^T(t-s,y,A)$ for all $0 \leq s < t$, $x \in \mathbb{R}^r$, and $A \in \mathbb{B}^r$
- $P\{x^T(t) \in A \mid x^T(s)\} = P^T(x^T(s),A)$ w.p.1 for all $0 \leq s < t$ and $A \in \mathbb{B}^r$

Under certain conditions (c.f. [31]), $P^T(\cdot,\cdot,\cdot)$ has an invariant Gibbs measure $\Pi^T(\cdot)$, i.e.,

$$\Pi^T(A) = \int \Pi^T(dx)\, P^T(t,x,A) \qquad \forall\, t \geq 0\,, \quad \forall\, A \in \mathbb{B}^r\,,$$

where

$$\Pi^T(A) = \frac{\int_A \exp(-U(x)/T)\, dx}{\int \exp(-U(y)/T)\, dy} \qquad \forall\, A \in \mathbb{B}^r\,,$$

and furthermore

$$P\{x^T(t) \in \cdot\} \rightarrow \Pi^T(\cdot) \quad \text{weakly} \quad \text{as} \quad t \rightarrow \infty\,. \tag{4.2}$$

Now for suitable $U(\cdot)$

$$\Pi^T(\cdot) \rightarrow \Pi^*(\cdot) \quad \text{weakly} \quad \text{as} \quad T \rightarrow 0 \tag{4.3}$$

where $\Pi^*(\cdot)$ is a probability measure on $(\mathbb{R}^r,\mathbb{B}^r)$ with support in the set S of global minima of $U(\cdot)$; see [17] for conditions under which (4.3) holds and a characterization of $\Pi^*(\cdot)$ in terms of the Hessian of $U(\cdot)$. In view of (4.2) and (4.3) the idea behind the Langevin algorithm is that by choosing $T = T(t) \rightarrow 0$ slowly enough hopefully

$$P\{x(t) \in \cdot\} \approx \Pi^{T(t)}(\cdot) \qquad (t \text{ large})$$

and then perhaps

$$P\{x(t) \in \cdot\} \rightarrow \Pi^*(\cdot) \quad \text{weakly} \quad \text{as} \quad t \rightarrow \infty \tag{4.4}$$

and consequently $x(t)$ converges to S in probability.

The Langevin and the annealing algorithms both have a stochastic descent behavior whereby "downhill" moves are modified probabilistically by "uphill" moves with fewer and fewer uphill moves as time tends to infinity and temperature tends to zero. However, the simulations of these Monte Carlo algorithms are quite different. To simulate sample paths of $x(\cdot)$ we might discretize (in time) the Langevin algorithm as

$$x_{k+1}^{\epsilon} = x_k^{\epsilon} - \nabla U(x_k^{\epsilon})\epsilon + \sqrt{2T(k\epsilon)\epsilon}\, w_k \,, \tag{4.5}$$

where $\{w_k\}$ is a sequence of standard $\mathbb{R}^r$-valued Gaussian random variables and ϵ is a (positive) discretization interval, and simulate sample paths of $\{x_k^{\epsilon}\}$ by generating pseudorandom Gaussian variates. $\nabla U(\cdot)$ may be computed from an analytical formula or approximated in a standard fashion. Compare this simulation with that of the annealing algorithm (see Chapter 2).

Geman reports some encouraging numerical results have been obtained by Aluffi-Pentini et. al. [32] with a modified Langevin algorithm which uses an interactive temperature schedule. Tests have been run on $U(\cdot)$ defined on $\mathbb{R}^r$ with $r = 1,\ldots,14$. Gidas also reports a numerical experiment with a single $U(\cdot)$ defined on $\mathbb{R}$ with 400 local minima. He suggests that a combination of the Langevin algorithm with the popular multistart technique (c.f. [29]) might improve the performance obtained by using either approach alone. We remark here that comparing different global optimization algorithms is in general a very difficult problem. Rubenstein [29] discusses some analytical methods for comparing different algorithms. Dixon and Szego [5] have attempted to define a standard set of test functions which might be used to empirically compare different algorithms. It is not clear that either of these methods are suitable for evaluating the performance of the Langevin algorithm. These tools it seems were designed to compare algorithms which in some way take advantage of the structure of smooth functions on low dimensional spaces. We regard the Langevin algorithm as a "universal" algorithm which may be used on functions defined on high dimensional space whose structure is essentially unknown or cannot be simply characterized. It seems that the best test for the Langevin algorithm is the particular problem one wishes to solve.

We shall now outline those convergence results for the Langevin algorithm which are known to us. We refer the reader to the specific paper for full details.

Geman and Hwang [9] were the first to obtain a convergence result for the Langevin algorithm. They consider a version of the Langevin algorithm restricted to a compact subset of $\mathbb{R}^r$ (using reflection barriers). They show that for a temperature schedule of the form

$$T(t) = \frac{c}{\log t} \qquad (t \text{ large})$$

that if c is no smaller than the difference between the maximum and minimum values of $U(\cdot)$ then (4.4) is obtained.

Gidas [11] has obtained necessary and sufficient conditions for the convergence of the Langevin algorithm in all of $\mathbb{R}^r$, using partial differential equation methods. He shows that there exists a constant Δ^* such that for temperature schedules $T(t)\downarrow 0$, (4.4) holds iff

$$\int_0^\infty \exp\left(-\frac{\Delta^*}{T(t)}\right) dt = \infty$$

Furthermore, the constant Δ^* is the natural continuous analog of Hajek's constant (see (2.10)). Chiang et. al. [4] have also obtained sufficient conditions for the convergence of the Langevin algorithm in all of $\mathbb{R}^r$ using large deviations theory.

Kushner [21] has obtained a detailed picture of the asymptotic behavior of a class of diffusions related to the Langevin algorithm and certain discrete-time approximations as well. Kushner considers (in discrete-time) an algorithm of the form

$$X_{k+1} = X_k + a_k b(X_k,\xi_k) + \sqrt{2}\, a_k \sigma(X_k) w_k \tag{4.6}$$

where $\{\xi_k\}$ is a sequence of bounded random variables and

$$a_k = \frac{c}{\log k} \qquad (k \text{ large}) .$$

In the special case where $\bar{b}(\cdot) = E\{b(\cdot,\xi_k)\} = -\nabla U(\cdot)$ and $\sigma(\cdot) = I$, (4.6) is a stochastic approximation version of the Langevin algorithm with noisy measurements of $\nabla U(\cdot)$. We shall refer to the Monte Carlo algorithm which simulates the sample paths of $\{X_k\}$ as *Kushner's algorithm.*

We remark that the conditions under which the above results are obtained typically include

(i) $U(\cdot)$ has continuous second-order partial derivatives

(ii) The local minima of $U(\cdot)$ consist of a finite number of compact sets; for Gidas' result it is actually required that the local minima be isolated and nondegenerate.

These assumptions are stronger than these assumed in Theorem 3.4, where it was only required that $U(\cdot)$ be continuous on a compact metric space. Of course the conclusion of Theorem 3.4 is only that the annealing algorithm visits a given neighborhood of S infinitely often with probability 1, whereas the above results show convergence of the Langevin algorithm to S in probability.

In this Chapter we shall examine certain issues concerning the Langevin and annealing algorithms which seem important to us and apparently have not been considered elsewhere. We proceed as follows. We have seen that the motivation behind the annealing and Langevin algorithms is quite similar. The first question we would like to answer is:

- what more can be said about the relationship between the annealing and Langevin algorithms?

In 4.2 we shall show that an annealing chain driven by white Gaussian noise converges in a certain sense to a Langevin diffusion. Now it seems clear that the annealing algorithm and the Langevin algorithm each have certain advantages. The Langevin algorithm, for example, looks like (for large time and small temperature) a gradient descent algorithm, and gradient descent algorithms and their higher order generalizations such as Newton's algorithm, which are "local" algorithms in the sense that they use only the value of the objective function and a finite number of derivatives at the current iterate to obtain the next iterate, are efficient at finding local minima. The annealing algorithm, on the other hand, is not strictly "local" in that it uses the value of the objective function in some set containing the current iterate to obtain the next iterate. In this sense, the annealing algorithm might be called "semilocal" or even "global" depending on how much of the objective function is used. Following the usual thinking behind both the annealing and Langevin algorithms, the idea is to make large fluctuations initially and small descent-like moves eventually. In view of these considerations, the second question we would like to answer is:

- is there a natural hybrid algorithm whose initial behavior resembles the annealing algorithm an whose large time behavior is similar to the Langevin algorithm?

In 4.3 we propose such an algorithm based on the results of 4.2.

4.2 Convergence of the Annealing Chain to a Langevin Diffusion

In this Section we shall examine the relationship between the annealing and Langevin algorithms. We shall show using a result of Kushner's [22] on the weak convergence of interpolated Markov chains to diffusions that a parameterized family of annealing chains driven by white Gaussian noise interpolated into piecewise constant processes converge weakly to a time-scaled solution of the Langevin equation. The weak convergence here is in the sense that the probability measures induced by the interpolated chains on the path space of functions without discontinuities of the second kind

converge weakly to the probability measure induced by the limit diffusion. This technique is the same one used to justify the popular diffusion approximation method, whereby a complicated possibly non-Markovian process is approximated by a simpler diffusion process (c.f. [23]).

Let $D^r[0,\overline{T}]$ denote the space of $\mathbb{R}^r$-valued càdlàg functions on $[0,\overline{T}]$ with $0 < \overline{T} < \infty$, i.e., functions which are right-continuous on $[0,\overline{T})$, have left-hand limits on $(0,\overline{T}]$, and are left continuous at $\overline{T}$. The following elementary results on weak convergence of probability measures may be found in [2]. There is a metric $d_T(\cdot,\cdot)$ on $D^r[0,\overline{T}]$ with respect to which $D^r[0,\overline{T}]$ is a complete separable metric space, and if $f(\cdot)\in D^r[0,T]$ and $\{f_n(\cdot)\}$ is a sequence in $D^r[0,\overline{T}]$ then the convergence of $f_n(\cdot)$ to $f(\cdot)$ in $D^r[0,\overline{T}]$ implies convergence at all points of continuity of $f(\cdot)$ (convergence of $f_n(\cdot)$ to $f(\cdot)$ in $D^r[0,\overline{T}]$ is roughly equivalent to uniform convergence outside of any neighborhood of the discontinuity points of $f(\cdot)$). Let $\xi(\cdot)$, $\{\xi_\epsilon(\cdot) : \epsilon > 0\}$ be processes with sample paths in $D^r[0,\overline{T}]$, or equivalently, random variables which take values in $D^r[0,\overline{T}]$, and let $\mu(\cdot)$, $\{\mu_\epsilon(\cdot) : \epsilon > 0\}$ be the probability measures they induce on the Borel subsets of $D^r[0,\overline{T}]$. We shall say that $\xi_\epsilon(\cdot)$ converges weakly to $\xi(\cdot)$ in $D^r[0,\overline{T}]$ and write $\xi_\epsilon(\cdot)\rightarrow\xi(\cdot)$ weakly (in $D^r[0,\overline{T}]$) if $\mu_\epsilon(\cdot)$ converges weakly to $\mu(\cdot)$ as $\epsilon\rightarrow 0$, i.e., if

$$\lim_{\epsilon\rightarrow 0} \int f(x)\, d\mu_\epsilon(x) = \int f(x)\, d\mu(x)$$

for all bounded continuous $f(\cdot)$ on $D^r[0,\overline{T}]$. Let $D^r[0,\infty)$ denote the set of $\mathbb{R}^r$-valued functions on $[0,\infty)$ which are right-continuous on $[0,\infty)$ and have left-hand limits on $(0,\infty)$. Let

$$d(f,g) = \sum_{n=1}^{\infty} \frac{1}{2^n}\, d_n(f,g) \qquad \forall\, f,g\in D^r[0,\infty) .$$

$d(\cdot,\cdot)$ is a metric on $D^r[0,\infty)$ with respect to which $D^r[0,\infty)$ is a complete separable metric space, and we can define the weak convergence of processes with sample paths in $D^r[0,\infty)$ analogously to $D^r[0,\overline{T}]$ with $\overline{T}$ finite.

Suppose $\xi_\epsilon(\cdot)\rightarrow\xi(\cdot)$ weakly (in $D^r[0,\overline{T}]$) as $\epsilon\rightarrow 0$ with $0 < \overline{T} \leq \infty$. Then it can be shown that the set of points $t\in[0,\overline{T}]$ such that $\mu(\{\xi(t_-) \neq \xi(t)\}) > 0$ is at most countable. Let

$$C = \{t\in[0,\overline{T}] : \mu(\{\xi(t_-) \neq \xi(t)\}) = 0\} .$$

Then it can also be shown that for any points $t_1,\ldots,t_k\in C$ the multivariate distributions of $\{\xi_\epsilon(t_1),\ldots,\xi_\epsilon(t_k)\}$ converge to the multivariate distributions of $\{\xi(t_1),\ldots,\xi(t_k)\}$ as $\epsilon\rightarrow 0$. But the weak convergence of $\xi_\epsilon(\cdot)$ to $\xi(\cdot)$ says much more than this: if $f(\cdot)$ is a continuous functional on $D^r[0,\overline{T}]$ (or just μ-a.s.

continuous) then $f(\xi_\epsilon(\cdot)) \longrightarrow f(\xi(\cdot))$ weakly as $\epsilon \longrightarrow 0$.

Let $C^r[0,\bar{T}]$ denote the space of $\mathbb{R}^r$-valued continuous functions on $[0,\bar{T}]$ with $0 \leq T \leq \infty$. If we equip $C^r[0,\bar{T}]$ with the uniform topology for $T < \infty$ and with the topology of uniform convergence on compacts for $\bar{T} = \infty$, then $C^r[0,\bar{T}]$ is a complete separable metric space and we can define weak convergence of processes with sample paths in $C^r[0,\bar{T}]$. Our reason for using $D^r[0,\bar{T}]$ is simply that we shall make use of Kushner's result on the weak convergence of Markov chains interpolated into $D^r[0,\bar{T}]$. Kushner's stated reason for working with $D^r[0,\bar{T}]$ as opposed to $C^r[0,\bar{T}]$ is that it is easier to verify tightness (relative compactness) for a sequence of probability measures on the Borel subsets of $D^r[0,\bar{T}]$. If the limit process is a jump diffusion then of course it would be necessary to work with $D^r[0,\bar{T}]$, but this is not an issue here since our limit processes are assumed to be ordinary (continuous sample paths with probability one) diffusions.

We now set up the notation necessary to state Kushner's Theorem on the weak convergence of interpolated Markov chains. It will be notationally convenient in the sequel to assume that all processes are defined on a common probability space (Ω, F, P) and we shall do so without further comment. Let $0 < \bar{T} < \infty$. Let $F(\cdot,\cdot)$ and $F_\epsilon(\cdot,\cdot)$, $\epsilon > 0$, be $\mathbb{R}^r$-valued Borel functions on $\mathbb{R}^r \times [0,\bar{T}]$, and let $G(\cdot,\cdot)$ and $G_\epsilon(\cdot,\cdot)$, $\epsilon > 0$, be $r \times r$ matrix-valued Borel functions on $\mathbb{R}^r \times [0,\bar{T}]$. For each $\epsilon > 0$ let $\{\xi_k^\epsilon\}$ be a Markov chain with state-space $\mathbb{R}^r$ such that

$$E\{\xi_{k+1}^\epsilon - \xi_k^\epsilon | \xi_k^\epsilon\} = F_\epsilon(\xi_k^\epsilon, k\epsilon)\epsilon ,$$
$$E\{(\xi_{k+1}^\epsilon - \xi_k^\epsilon) \otimes (\xi_{k+1}^\epsilon - \xi_k^\epsilon) | \xi_k^\epsilon\} = G_\epsilon(\xi_k^\epsilon, k\epsilon)\, G_\epsilon'(\xi_k^\epsilon, k\epsilon)\epsilon ,$$

with probability one. Interpolate $\{\xi_k^\epsilon\}$ into a process $\xi_\epsilon(\cdot)$ with sample paths in $D^r[0,\bar{T}]$ by

$$\xi_\epsilon(t) = \xi_k^\epsilon \qquad \forall\ (k-1)\epsilon \leq t < k\epsilon , \qquad \forall\ k = 1,\ldots, \left\lceil \frac{\bar{T}}{\epsilon} \right\rceil .$$

Here is Kushner's Theorem in slightly modified form.

Theorem 4.1 (Kushner [22]). Assume

(K1) $F(\cdot,\cdot)$, $G(\cdot,\cdot)$ are bounded and continuous

(K2) $F_\epsilon(\cdot,\cdot)$, $G_\epsilon(\cdot,\cdot)$ are uniformly bounded for small $\epsilon > 0$

$$\text{(K3)}\quad \mathrm{E}\left\{\sum_{k=1}^{\lfloor \bar{T}/\epsilon \rfloor}\left[|F_\epsilon(\xi_k^\epsilon,k\epsilon)-F(\xi_k^\epsilon,k\epsilon)|^2 + |G_\epsilon(\xi_k^\epsilon,k\epsilon)-G(\xi_k^\epsilon,k\epsilon)|^2\right]\epsilon\right\} \to 0$$

as $\epsilon \to 0$

$$\text{(K4)}\quad \mathrm{E}\left\{\sum_{k=1}^{\lfloor \bar{T}/\epsilon \rfloor}\left[|\xi_{k+1}^\epsilon-\xi_k^\epsilon-F_\epsilon(\xi_k^\epsilon,k\epsilon)\epsilon|^{2+\alpha}\right\} \to 0$$

as $\epsilon \to 0$ for some $\alpha > 0$.

Let $v(\cdot)$ be a standard r-dimensional Wiener process and assume that

$$d\xi(t) = F(\xi(t),t)dt + G(\xi(t),t)dv(t), \qquad 0 \le t \le \bar{T},$$

has a unique solution $\xi(\cdot)$ (in the sense of multivariate distributions) with initial condition $\xi(0) = \xi_0$. Assume that

$$\xi_1^\epsilon \to \xi_0 \quad \text{weakly as } \epsilon \to 0.$$

Then

$$\xi_\epsilon(\cdot) \to \xi(\cdot) \quad \text{weakly } (\text{in } D^r[0,\bar{T}]) \text{ as } \epsilon \to 0.$$

Consider now the following family of Markov chains. Let $U(\cdot)$ and $T(\cdot)$ be defined as in 4.1. For each $\epsilon > 0$ let $\{z_k^\epsilon\}$ be a Markov chain with state space $\mathbb{R}^r$ and 1-step transition functions $\{P_k^\epsilon(\cdot,\cdot)\}$ given by†

$$P_k^\epsilon(x,A) = \int_A s_k^\epsilon(x,y)\, dN(x,\epsilon I)(y) + \gamma_k^\epsilon(x)\, \delta(x,A) \tag{4.7}$$

for all $x \in \mathbb{R}^r$ and $A \in \mathcal{B}^r$, where

†see Chapter 3 for general state space Markov chain notation

$$s_k^\epsilon(x,y) = \begin{cases} \exp\left[-\dfrac{U(y)-U(x)}{T(k\epsilon)}\right] & \text{if } U(y) > U(x) \\ 1 & \text{if } U(y) \le U(x)\,, \end{cases} \tag{4.8}$$

$$\gamma_k^\epsilon(x) = 1 - \int s_k^\epsilon(x,y)\, dN(x,\epsilon I)\,(y)\,, \tag{4.9}$$

and $\delta(x,\cdot)$ is the unit measure concentrated at x, for all $x,y\in\mathbb{R}^r$. Comparing (4.7) and (3.2) it is seen that $\{z_k^\epsilon\}$ is infact an annealing chain of the type introduced in Chapter 3 with state space the measure space (Σ,B,ϕ) where $\Sigma = \mathbb{R}^r$, $B = \mathbb{B}^r$, $\phi(\cdot)$ is Lebesque measure, and

$$Q(x,A) = \int_A q(x,y)\,\phi(dy) = N(x,\epsilon I)\,(A) \qquad \forall\, A\in\mathbb{B}^r$$

(hence the annealing chain is "driven" by white Gaussian noise). It will be convenient to introduce the following notation. For each $\epsilon > 0$ let

$$s(x,y,t) = \begin{cases} \exp\left[-\dfrac{U(y)-U(x)}{T(t)}\right] & \text{if } U(y) > U(x) \\ 1 & \text{if } U(y) \le U(x)\,, \end{cases}$$

$$\gamma_\epsilon(x,t) = 1 - \int s(x,y,t)\, dN(x,\epsilon I)\,(y)\,,$$

for all $x,y\in\mathbb{R}^r$ and $t \ge 0$, and let

$$P_\epsilon(x,A,t) = \int_A s(x,y,t)\, dN(x,\epsilon I)\,(y) + \gamma_\epsilon(x,t)\,\delta(x,A)$$

for all $x\in\mathbb{R}^r$, $A\in\mathbb{B}^r$, and $t \ge 0$. Then

$$P_\epsilon(x,A,k\epsilon) = P_k^\epsilon(x,A) \qquad \forall\, x\in\mathbb{R}^r\,,\ \forall\, A\in\mathbb{B}^r\,.$$

For each $\epsilon > 0$, $x\in\mathbb{R}^r$ and $t \ge 0$ let

$$b_\epsilon(x,t) = \frac{1}{\epsilon}\int (y-x)\, P_\epsilon(x,dy,t)\,,$$

$$a_\epsilon(x,t) = \frac{1}{\epsilon}\int (y-x)\otimes(y-x)\, P_\epsilon(x,dy,t)\,,$$

and $\sigma_\epsilon(x,t)$ be a positive square root of $a_\epsilon(x,t)$ i.e.

$$\sigma_\epsilon(x,t)\sigma_\epsilon'(x,t) = a_\epsilon(x,t)\,.$$

Since $P_\epsilon(\cdot,\cdot,k\epsilon) = P_k^\epsilon(\cdot,\cdot)$ is a (regular wide-sense) conditional distribution for

z_{k+1}^{ϵ} given z_k^{ϵ},

$$E\{z_{k+1}^{\epsilon} - z_k^{\epsilon} | z_k^{\epsilon}\} = b_{\epsilon}(z_k^{\epsilon}, k\epsilon)\epsilon$$
$$E\{(z_{k+1}^{\epsilon} - z_k^{\epsilon}) \otimes (z_{k+1}^{\epsilon} - z_k^{\epsilon}) | z_k^{\epsilon}\} = \sigma_{\epsilon}(z_k^{\epsilon}, k\epsilon)\sigma_{\epsilon}'(z_k^{\epsilon}, k\epsilon)\epsilon$$

with probability one. Interpolate $\{z_k^{\epsilon}\}$ into $z_{\epsilon}(\cdot)$ with sample paths in $D^r[0,\overline{T}]$ by

$$z_{\epsilon}(t) = z_k^{\epsilon} \qquad \forall\ (k-1)\epsilon \leq t < k\epsilon\ , \qquad \forall\ k = 1,\ldots, \left\lceil \frac{\overline{T}}{\epsilon} \right\rceil .$$

Here is our convergence theorem.

Theorem 4.2 Assume

(A1) $U(\cdot)$ is continuously differentiable, $\nabla U(\cdot)$ is bounded and Lipshitz

(A2) $T(\cdot)$ is continuous

Let $w(\cdot)$ be a standard r-dimensional Wiener process, and let $z(\cdot)$ be a solution of†

$$dz(t) = -\frac{\nabla U(z(t))}{2T(t)}\, dt + dw(t)\ , \qquad 0 \leq t \leq \overline{T}\ , \tag{4.10}$$

with initial condition $z(0) = z_0$. Assume that

$$z_1^{\epsilon} \longrightarrow z_0 \quad \text{weakly} \quad \text{as} \quad \epsilon \longrightarrow 0\ .$$

Then

$$z_{\epsilon}(\cdot) \longrightarrow z(\cdot) \quad \text{weakly} \quad (\text{in } D^r[0,\overline{T}]) \quad \text{as} \quad \epsilon \longrightarrow 0\ .$$

Remark Let $\tau(\cdot)$ be a solution of

$$\dot{\tau}(t) = 2T(\tau(t))\ , \qquad 0 \leq t \leq \overline{T}\ ,$$

and let

$$\tilde{z}(t) = z(\tau(t))\ , \qquad \tilde{T}(t) = T(\tau(t))\ , \qquad 0 \leq t \leq \overline{T}\ .$$

Clearly, $\tilde{z}(\cdot)$ is a Markov process with continuous sample paths with probability one. Now by standard calculations

†it is well known that (4.10) has a (strongly) unique solution under (A1), (A2)

$$E\{\tilde{z}(t+h) - \tilde{z}(t)|\tilde{z}(t)\} = -\nabla U(\tilde{z}(t))h + O(h^{3/2})$$

$$E\{(\tilde{z}(t+h) - \tilde{z}(t))\otimes(\tilde{z}(t+h) - \tilde{z}(t))|\tilde{z}(t)\} = \int_t^{t+h} \sqrt{2\tilde{T}(s)}\, ds\cdot I + O(h^2)$$

as $h\to 0$, uniformly for $0 \le t \le \bar{T}$, with probability one. Hence by a Theorem of Doob's [6, p. 288] there exists a standard r-dimensional Wiener process $\tilde{w}(\cdot)$ such that $\tilde{z}(\cdot)$ is the solution of

$$d\tilde{z}(t) = -\nabla U(\tilde{z}(t))dt + \sqrt{2\tilde{T}(t)}\, d\tilde{w}(t)\,, \qquad 0 \le t \le \bar{T}\,. \tag{4.11}$$

Hence the interpolated annealing chain $z_\epsilon(\cdot)$ converges weakly to $z(\cdot)$, which is infact a time-scaled solution of the Langevin equation (4.11).

We shall need several lemmas before we can apply Theorem 4.1 to prove Theorem 4.2. Let

$$\hat{s}(x,y,t) = \begin{cases} \exp\left[-\dfrac{(\nabla U(x),\, y-x)}{T(t)}\right] & \text{if } (\nabla U(x),\, y-x) > 0 \\ 1 & \text{if } (\nabla U(x),\, y-x) \le 0 \end{cases} \tag{4.12}$$

for all $x,y\in\mathbb{R}^r$ and $t \ge 0$.

Lemma 4.1 Assume (A1), (A2). Then there exists a constant K such that

$$|s(x,y,t) - \hat{s}(x,y,t)| \le K|y-x|^2\,, \qquad \forall\, x,y\in\mathbb{R}^r\,, \qquad \forall\, 0 \le t \le \bar{T}\,.$$

Proof Let

$$f(x,y) = U(y) - U(x) - (\nabla U(x),\, y-x) \qquad \forall\, x,y\in\mathbb{R}^r\,.$$

By the Mean Value Theorem and (A1) there exists a constant c such that

$$|f(x,y)| \le c|y-x|^2 \qquad \forall\, x,y\in\mathbb{R}^r\,,$$

and by (A2) there exists a constant K such that

$$\frac{|f(x,y)|}{T(t)} \le K|y-x|^2 \qquad \forall\, x,y\in\mathbb{R}^r\,,\ \forall\, 0 \le t \le \bar{T}\,.$$

Suppose $U(y) - U(x) \ge 0$ and $(\nabla U(x),\, y-x) \le 0$. Then $|U(y) - U(x)| \le |f(x,y)|$ and since $|1 - e^x| \le |x|$ for $x \le 0$

$$|s(x,y,t) - \hat{s}(x,y,t)| = \left|1 - \exp\left[-\frac{U(y) - U(x)}{T(t)}\right]\right|$$

$$\leq \frac{|U(y) - U(x)|}{T(t)}$$

$$\leq \frac{|f(x,y)|}{T(t)}$$

$$\leq K \cdot |y-x|^2 \qquad \forall\, x,y \in \mathbb{R}^r\,,\ \forall\, 0 \leq t \leq \bar{T}\,.$$

The same inequality holds if $U(y) - U(x) \leq 0$ and $(\nabla U(x),\, y-x) \geq 0$. Suppose that $U(y) - U(x) \geq 0$ and $(\nabla U(x),\, y-x) \geq 0$. Then

$$|s(x,y,t) - \hat{s}(x,y,t)| \leq \left|1 - \exp\left[-\frac{|f(x,y)|}{T(t)}\right]\right|$$

$$\leq \frac{|f(x,y)|}{T(t)}$$

$$\leq K \cdot |y-x|^2 \qquad \forall\, x,y \in \mathbb{R}^r\,,\ \forall\, 0 \leq t \leq \bar{T}\,.$$

The Lemma follows by combining the various cases. □

The following two Lemmas provide the crucial estimates of the local drift $b_\epsilon(\cdot,\cdot)$ and local covariance $a_\epsilon(\cdot,\cdot)$ of $z_\epsilon(\cdot)$. The simple estimate

$$\int |y|^n \, dN(0,\epsilon I)\,(y) = O(\epsilon^{n/2}) \qquad \text{as} \quad \epsilon \longrightarrow 0$$

for $n \in \mathbb{N}$ will be used frequently in the sequel.

Lemma 4.2 Assume (A1), (A2). Then

$$b_\epsilon(x,t) = -\frac{\nabla U(x(t))}{2T(t)} + O(\epsilon^{1/2}) \qquad \text{as} \quad \epsilon \longrightarrow 0\,,$$

uniformly for $x \in \mathbb{R}^r$, $0 \leq t \leq \bar{T}$.

Proof By Lemma 4.1 there exists a constant K such that

$$|s(x,y,t) - \hat{s}(x,y,t)| \leq K\,|y-x|^2 \qquad \forall\, x,y \in \mathbb{R}^r\,,\ \forall\, 0 \leq t \leq \bar{T}\,.$$

Hence

$$b_\epsilon(x,t) = \frac{1}{\epsilon} \int (y-x)\, P_\epsilon(x,dy,t)$$

$$= \frac{1}{\epsilon} \int (y-x)\, s(x,y,t)\, dN(x,\epsilon I)\,(y)$$

$$= \frac{1}{\epsilon} \int (y-x)\, \hat{s}(x,y,t)\, dN(x,\epsilon I)\,(y)$$

$$+ \frac{1}{\epsilon} \int (y-x) \left[s(x,y,t) - \hat{s}(x,y,t) \right] dN(x,\epsilon I)\,(y)$$

$$= \frac{1}{\epsilon} \int (y-x)\, \hat{s}(x,y,t)\, dN(x,\epsilon I)\,(y) + O(\epsilon^{1/2}) \quad \text{as} \quad \epsilon \to 0\,,$$

uniformly for $x \in \mathbb{R}^r$ and $0 \le t \le \bar{T}$. Substituting for $\hat{s}(\cdot,\cdot,\cdot)$ from (4.12) gives

$$b_\epsilon(x,t) = \frac{1}{\epsilon^{1/2}} \int_{(y,\nabla U(x))\le 0} y\, dN(0,I)\,(y)$$

$$+ \frac{1}{\epsilon^{1/2}} \int_{(y,\nabla U(x))>0} y \exp\left[-\frac{(\nabla U(x),y)}{T(t)}\, \epsilon^{1/2} \right] dN(0,I)\,(y) + O(\epsilon^{1/2}) \qquad (4.13)$$

as $\epsilon \to 0$, uniformly for $x \in \mathbb{R}^r$ and $0 \le t \le \bar{T}$. Clearly,

$$b_\epsilon(x,t) = -\frac{\nabla U(x)}{2T(t)} + O(\epsilon^{1/2}) \qquad \text{as} \quad \epsilon \to 0$$

uniformly on $\{x \in \mathbb{R}^r : \nabla U(x) = 0\} \times [0,\bar{T}]$. Hence we may assume that $\nabla U(x) \neq 0$ for all $x \in \mathbb{R}^r$. Let

$$\alpha(x,t) = \frac{1}{2} \left[\frac{|\nabla U(x)|}{T(t)} \right]^2, \quad \beta(x,t) = \frac{|\nabla U(x)|^2}{T(t)}\,, \qquad (4.14)$$

for all $x \in \mathbb{R}^r$ and $t \ge 0$. By (A1), (A2) $\alpha(\cdot,\cdot)$ and $\beta(\cdot,\cdot)$ are bounded on $\mathbb{R}^r \times [0,\bar{T}]$. Now completing the square in the second integrand in (4.13) gives

$$b_\epsilon(x,t) = \frac{1}{\epsilon^{1/2}} \int_{(y,\nabla U(x))\le 0} y\, dN(0,I)\,(y)$$

$$+ \frac{1}{\epsilon^{1/2}} \int_{(y,\nabla U(x))\ge 0} y \exp(\alpha(x,t)\epsilon)\, dN\left(-\frac{\nabla U(x)}{T(t)}\,\epsilon^{1/2}, I\right)(y) + O(\epsilon^{1/2})$$

$$= \frac{1}{\epsilon^{1/2}} \int_{(y,\nabla U(x))\le 0} y\, dN(0,I)\,(y) + \frac{1}{\epsilon^{1/2}} \int_{(y,\nabla U(x))\ge \beta(x,t)\epsilon^{1/2}} y\, dN(0,I)\,(y)$$

$$- \frac{\nabla U(x)}{T(t)} N(0,I)\,\{y:\ (y,\nabla U(x)) \ge \beta(x,t)\epsilon^{1/2}\} + O(\epsilon^{1/2})$$

$$= -\frac{\nabla U(x)}{T(t)} f(\nabla U(x), O(\epsilon^{1/2}))$$

$$+ \frac{1}{\epsilon^{1/2}} [g(\nabla U(x), O(\epsilon^{1/2})) - g(\nabla U(x),0)] + O(\epsilon^{1/2}), \qquad (4.16)$$

as $\epsilon \to 0$, uniformly for $x \in \mathbb{R}^r$ and $0 \le t \le \bar{T}$, where

$$f(u,\delta) = N(0,I)\,\{y:\ (y,u) \ge |u|\delta\},$$
$$g(u,\delta) = \int_{(y,u)\ge|u|\delta} y\, dN(0,I)\,(y).$$

To proceed further we need to estimate $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$. We have

$$f(u,\delta) = N(0,I)\,\{y:\ (y,u) \ge |u|\delta\}$$
$$= N(0,1)\,[\delta,\infty)$$
$$= \frac{1}{2} - \int_0^\delta \frac{1}{(2\pi)^{1/2}} e^{-\xi^2/2}\, d\xi$$
$$= \frac{1}{2} + O(\delta) \qquad \text{as } \delta \to 0, \qquad (4.17)$$

uniformly for $u \in \mathbb{R}^r$. As for $g(\cdot,\cdot)$, we make the following

Claim

$$g(u,\delta) = \frac{e^{-\delta^2}}{(2\pi)^{1/2}} \frac{u}{|u|} \qquad (4.18)$$

for all $\delta > 0$ and $u \in \mathbb{R}^r \setminus \{0\}$.

Suppose the Claim is true. Then combining (4.16)-(4.18) gives

$$b_\epsilon(x,t) = -\frac{\nabla U(x)}{2T(t)} + \frac{e^{O(\epsilon)} - 1}{(2\pi\epsilon)^{1/2}} \frac{\nabla U(x)}{|\nabla U(x)|} + O(\epsilon^{1/2})$$
$$= -\frac{\nabla U(x)}{2T(t)} + O(\epsilon^{1/2}) \qquad \text{as } \epsilon \to 0 ,$$

uniformly for $x \in \mathbb{R}^r$ and $0 \le t \le \bar{T}$, and we obtain the Theorem. It remains to prove the Claim.

Proof of Claim Fix $u \in \mathbb{R}^r \setminus \{0\}$, let

$$n_1 = \frac{u}{|u|} ,$$

and extend n_1 to an orthonormal basis $\{n_1,...,n_r\}$ for $\mathbb{R}^r$. Also let $\{e_1,...,e_r\}$ be the standard basis for $\mathbb{R}^r$, and $L(\cdot)$ be the (orthogonal) linear mapping from $\mathbb{R}^r$ into $\mathbb{R}^r$ such that $L(e_i) = n_i$ for all $i = 1,...,r$. Applying the change of variable formula and using the fact that $L(\cdot)$ is an isometry and the adjoint $L^*(\cdot) = L^{-1}(\cdot)$ gives

$$g(u,\delta) = \int_{(y,u)\ge|u|\delta} y \, dN(0,I)\,(y)$$
$$= \int_{(y,n_1)\ge\delta} \sum_{i=1}^{r} (y,n_i) n_i \cdot dN(0,I)\,(y)$$
$$= \sum_{i=1}^{r} n_i \int_{(Lz,n_i)\ge\delta} (Lz,n_i)\, dN(0,I)\,(z)$$
$$= \sum_{i=1}^{r} n_i \int_{(z,L^* n_i)\ge\delta} (z,L^* n_i)\, dN(0,I)\,(z)$$
$$= \sum_{i=1}^{r} n_i \int_{(z,e_i)\ge\delta} (z,e_i)\, dN(0,I)\,(z)$$
$$= n_1 \int_{\delta}^{\infty} \xi \, dN(0,1)\,(\xi)$$
$$= n_1 \int_{\delta}^{\infty} \xi \cdot \frac{1}{(2\pi)^{1/2}} \exp\left(-\xi^2/2\right) d\xi$$
$$= \frac{e^{-\delta^2}}{(2\pi)^{1/2}} \frac{u}{|u|} \qquad \forall\, \delta > 0 .$$

This completes the proof of the Claim and hence the Theorem. □

Lemma 4.3 Assume (A1), (A2). Then

$$a_\epsilon(x,t) = I + O(\epsilon^{1/2}) \qquad \text{as} \quad \epsilon \to 0$$

uniformly for $x \in \mathbb{R}^r$ and $0 \le t \le \overline{T}$.

Proof Proceeding as in the proof the Lemma 4.2

$$a_\epsilon(x,t) = \frac{1}{\epsilon} \int (y-x) \otimes (y-x)\, P_\epsilon(x,dy,t)$$

$$= \frac{1}{\epsilon} \int (y-x) \otimes (y-x)\, \hat{s}(x,y,t)\, dN(x,\epsilon I)\,(y) + O(\epsilon)$$

$$= \int_{(y,\nabla U(x)) \le 0} y \otimes y\, dN(0,I)\,(y)$$

$$+ \int_{(y,\nabla U(x)) > 0} y \otimes y \exp\left[- \frac{(\nabla U(x),y)}{T(t)}\, \epsilon^{1/2} \right] dN(0,I)\,(y) + O(\epsilon) \qquad (4.19)$$

as $\epsilon \to 0$, uniformly for $x \in \mathbb{R}^r$ and $0 \le t \le \overline{T}$. Clearly,

$$a_\epsilon(x,t) = I + O(\epsilon) \quad \text{as} \quad \epsilon \to 0$$

uniformly on $\{x \in \mathbb{R}^r : \nabla U(x) = 0\} \times [0,\overline{T}]$. Hence we may assume that $\nabla U(x) \neq 0$ for all $x \in \mathbb{R}^r$. Let $\alpha(\cdot,\cdot)$, $\beta(\cdot,\cdot)$ be defined as in (4.14). Then completing the square in the second integrand in (4.19) gives

$$a_\epsilon(x,t) = \int_{(y,\nabla U(x)) \le 0} y \otimes y\, dN(0,I)\,(y)$$

$$+ \int_{(y,\nabla U(x)) \ge 0} y \otimes y \exp\left(\alpha(x,t)\epsilon\right) dN\left(- \frac{\nabla U(x)}{T(t)}\, \epsilon^{1/2}\, , I \right)(y) + O(\epsilon)$$

$$= \int_{(y,\nabla U(x)) \le 0} y \otimes y\, dN(0,I)\,(y) + \int_{(y,\nabla U(x)) \ge \beta(x,t)\epsilon^{1/2}} y \otimes y\, dN(0,I)\,(y) + O(\epsilon^{1/2})$$

$$= h(\nabla U(x),0) + h(\nabla U(x), O(\epsilon^{1/2})) + O(\epsilon^{1/2})\,, \qquad (4.20)$$

as $\epsilon \to 0$, uniformly for $x \in \mathbb{R}^r$ and $0 \le t \le \overline{T}$, where

$$h(u,\delta) = \int_{(y,u) \ge |u|\delta} y \otimes y\, dN(0,I)\,(y)\,.$$

To proceed further we need to estimate $h(\cdot,\cdot)$. We make the following

Claim

$$h(u,\delta) = \frac{1}{2} I + O(\delta) \qquad \text{as} \quad \delta \longrightarrow 0 \ , \ . \tag{4.21}$$

uniformly for $u \in \mathbb{R}^r$.

Suppose the Claim is true. Then combining (4.20) and (4.21) gives

$$a_\epsilon(x,t) = I + O(\epsilon^{1/2}) \qquad \text{as} \quad \epsilon \longrightarrow 0 \ ,$$

uniformly for $x \in \mathbb{R}^r$ and $0 \leq t \leq \overline{T}$, and we obtain the Theorem. It remains to prove the Claim.

Proof of Claim Fix $u \in \mathbb{R}^r \backslash \{0\}$ and let $\{n_1,\ldots,n_r\}$, $\{e_1,\ldots,e_r\}$, and $L(\cdot)$ be as in the proof of the Claim in Lemma 4.2. Then

$$\begin{aligned}
h(u,0) &= \int_{(y,u)\geq 0} y \otimes y \, dN(0,I)\,(y) \\
&= \int_{(y,n_1)\geq 0} \left(\sum_{i=1}^{r} (y,n_i) n_i \right) \otimes \left(\sum_{j=1}^{r} (y,n_j) n_j \right) dN(0,I)\,(y) \\
&= \sum_{i,j=1}^{r} n_i \otimes n_j \int_{(Lz,n_1)\geq 0} (Lz,n_i)\,(Lz,n_j)\, dN(0,I)\,(y) \\
&= \sum_{i,j=1}^{r} n_i \otimes n_j \int_{(z,L^* n_1)\geq 0} (z,L^* n_i)\,(z,L^* n_j)\, dN(0,I)\,(y) \\
&= \sum_{i,j=1}^{r} n_i \otimes n_j \int_{(z,e_1)\geq 0} (z,e_i)\,(z,e_j)\, dN(0,I)\,(y) \\
&= \sum_{i=1}^{r} n_i \otimes n_i \int_{(z,e_1)\geq 0} (z,e_i)^2 \, dN(0,I)\,(y) \\
&= n_1 \otimes n_1 \int_0^\infty \xi^2 \, dN(0,1)\,(\xi) + \sum_{i=2}^{r} n_i \otimes n_i \, N(0,1)\,(0,\infty) \\
&= \frac{1}{2} \sum_{i=1}^{r} n_i \otimes n_i \\
&= \frac{1}{2} I
\end{aligned} \tag{4.22}$$

Similarly,

$$
\begin{aligned}
h(u,0) - h(u,\delta) &= \int_{0\le(y,u)\le|u|\delta} y\otimes y \; dN(0,I) \\
&= \sum_{i-1}^{r} n_i\otimes n_i \int_{0\le(z,e_i)\le\delta} (z,e_i)^2 \; dN(0,I)\,(z) \\
&= n_1\otimes n_1 \int_0^{\delta} \xi^2 dN(0,1)\,(\xi) + \sum_{i-2}^{r} n_i\otimes n_i \; N(0,1)\,[0,\delta) \\
&= n_1\otimes n_1 \int_0^{\delta} \xi^2 \frac{1}{(2\pi)^{1/2}} \exp\,(-\xi^2/2)d\xi \\
&\quad + \sum_{i-2}^{r} n_i\otimes n_i \int_0^{\delta} \frac{1}{(2\pi)^{1/2}} \exp\,(-\xi^2/2)\,d\xi \\
&= n_1\otimes n_i\cdot O(\delta^3) + \sum_{i-2}^{r} n_i\otimes n_i \cdot O(\delta) \\
&= O(\delta) \qquad \text{as} \quad \delta\longrightarrow 0\,. \qquad (4.23)
\end{aligned}
$$

Combining (4.22) and (4.23) completes the proof of the Claim and hence the Theorem. □

Lemma 4.4 Assume (A1), (A2). Then

$$\sigma_\epsilon(x,t) = I + O(\epsilon^{1/2}) \qquad \text{as} \quad \epsilon\longrightarrow 0$$

uniformly for $x\in\mathbb{R}^r$ and $0 \le t \le \overline{T}$.

Proof By Lemma 4.3

$$a_\epsilon(x,t) = I + O(\epsilon^{1/2}) \qquad \text{as} \quad \epsilon\longrightarrow 0 \qquad (4.24)$$

uniformly for $x\in\mathbb{R}^r$ and $0 \le t \le \overline{T}$. Since $a_\epsilon(x,t)$ is self-adjoint, there exists an orthogonal matrix $L_\epsilon(x,t)$ such that

$$a_\epsilon(x,t) = L_\epsilon(x,t)\,\Lambda_\epsilon(x,t)\,L_\epsilon'(x,t) \qquad (4.25)$$

where

$$\Lambda_\epsilon(x,t) = \text{diag}\,(\lambda_{\epsilon,1}(x,t),\ldots,\lambda_{\epsilon,r}(x,t)) \qquad (4.26)$$

and the $\{\lambda_{\epsilon,i}(x,t) : i = 1,\ldots,r\}$ are the eigenvalues of $a_\epsilon(x,t)$, i.e., the solutions of $\det(\lambda I - a_\epsilon(x,t)) = 0$. Now if $A = [a_{ij}]$ is a real $r\times r$ matrix then det A may be expressed as

$$\det A = \sum_{p=\{p_i\}\in P} \operatorname{sgn}(p) \cdot a_{1p_1} \cdots a_{rp_r} \tag{4.27}$$

where P is the set of permutations of $\{1,\ldots,r\}$. Setting $A = \lambda I - a_\epsilon(x,t)$ and combining (4.24) and (4.27) gives

$$\det(\lambda I - a_\epsilon(x,t)) = (\lambda - 1)^r + (\lambda - 1)^{r-1}\, O(\epsilon^{1/2}) + \cdots + O(\epsilon^{r/2})$$

and so

$$|\lambda_{\epsilon,i}(x,t) - 1|^r = O(\max\{|\lambda_{\epsilon,i}(x,t) - 1|^{r-1}\, \epsilon^{1/2}\,,\ \epsilon^{r/2}\})$$

and consequently

$$\lambda_{\epsilon,i}(x,t) = 1 + O(\epsilon^{1/2})\,,$$

and since $(1+\delta)^{1/2} = 1 + O(\delta)$ as $\delta \to 0$,

$$\lambda_{\epsilon,i}^{1/2}(x,t) = 1 + O(\epsilon^{1/2}) \qquad \text{as } \epsilon \to 0\,, \tag{4.28}$$

uniformly for $x \in \mathbb{R}^r$ and $0 \le t \le \bar{T}$. Let

$$\sigma_\epsilon(x,t) = L_\epsilon(x,t)\, \Lambda_\epsilon^{1/2}(x,t)\, L_\epsilon'(x,t)\,.$$

Then by (4.25) $a_\epsilon(x,t) = \sigma_\epsilon(x,t)\, \sigma_\epsilon'(x,t)$, and by (4.26), (4.28), and the Schwartz inequality,

$$\sigma_\epsilon(x,t) = I + O(\epsilon^{1/2}) \qquad \text{as } \epsilon \to 0\,,$$

uniformly for $x \in \mathbb{R}^r$ and $0 \le t \le \bar{T}$, as required. □

Proof of Theorem 4.2 We shall apply Theorem 4.1 with $\xi_k^\epsilon = z_k^\epsilon$, $\xi_\epsilon(\cdot) = z_\epsilon(\cdot)$, $\xi(\cdot) = z(\cdot)$, and

$$F(x,t) = -\frac{\nabla U(x)}{2T(t)}\,, \qquad F_\epsilon(x,t) = b_\epsilon(x,t)$$
$$G(x,t) = I\,, \qquad G_\epsilon(x,t) = \sigma_\epsilon(x,t)\,.$$

In view of (A1), (A2) and Lemmas 4.2 and 4.4, (K1) and (K2) are satisfied. Now by Lemmas 4.2 and 4.4 there exists a constant c such that for small enough $\epsilon > 0$

$$\left|b_\epsilon(x,t) + \frac{\nabla U(x)}{2T(t)}\right| \le c\, \epsilon^{1/2}\,,$$

$$|\sigma_\epsilon(x,t) - I| \le c\, \epsilon^{1/2}\,,$$

for all $x \in \mathbb{R}^r$ and $0 \le t \le \bar{T}$. Hence

$$E\left\{\sum_{k=1}^{\lfloor \bar{T}/\epsilon \rfloor}\left[\left|b_\epsilon(z_k^\epsilon,k\epsilon)+\frac{\nabla U(z_k^\epsilon)}{2T(k\epsilon)}\right|^2+\left|\sigma_\epsilon(z_k^\epsilon,k\epsilon)-I\right|^2\right]\epsilon\right\}$$

$$\leq \sum_{k=1}^{\lfloor \bar{T}/\epsilon \rfloor} 2c^2\epsilon^2 \qquad (\epsilon \text{ small})$$

$$\leq 2c^2\bar{T}\epsilon \longrightarrow 0 \qquad \text{as } \epsilon\longrightarrow 0$$

and so (K3) is satisfied.

It remains to check (K4). Since $P_k^\epsilon(\cdot,\cdot)$ is a conditional distribution function for z_{k+1}^ϵ given z_k^ϵ we have that for every $n\in\mathbb{N}$

$$\begin{aligned}E\{|z_{k+1}^\epsilon - z_k^\epsilon|^n\} &= E\{E\{|z_{k+1}^\epsilon - z_k^\epsilon|^n|z_k^\epsilon\}\}\\ &= E\{\int |y - z_k^\epsilon|^n\, P_k^\epsilon(z_k^\epsilon,\, dy)\}\\ &\leq E\{\int |y|^n |dN(0,\epsilon I)\,(y)\}\\ &\leq c_n\,\epsilon^{n/2}\end{aligned}$$

for some constant c_n. Hence using the uniform boundedness of $b_\epsilon(\cdot,\cdot)$ for small ϵ

$$E\left\{\left|\sum_{k=1}^{\lfloor \bar{T}/\epsilon \rfloor} |z_{k+1}^\epsilon - z_k^\epsilon - b_\epsilon(z_k^\epsilon,k\epsilon)\epsilon|^4\right\}\leq \sum_{k=1}^{\lfloor \bar{T}/\epsilon \rfloor} d\cdot\epsilon^2 \qquad (\epsilon \text{ small})\right.$$

$$\leq d\bar{T}\cdot\epsilon\longrightarrow 0 \qquad \text{as } \epsilon\longrightarrow 0$$

for some constant d, and so (K4) is satisfied with $\alpha = 2$. The Theorem now follows from Theorem 4.1. □

4.3 Hybrid Annealing/Langevin Algorithm

In this Section we shall give a hybrid annealing/Langevin algorithm whose initial behavior resembles that of the annealing algorithm and whose large time behavior is similar to the Langevin algorithm. The development of this algorithm will be guided by Kushner's algorithm and the results of 4.2 on the relationship between the annealing algorithm and the Langevin algorithm. We note that the discussion in this Section is heuristic at points and more work need to be done.

We shall make use of the notation introduced in 4.1, 4.2. We shall assume that

$$T(t) = \frac{c}{\log t} \qquad (t \text{ large})$$

where c is a positive constant.

We start by considering Kushner's algorithm (4.6) with $b(x,\xi) = -\nabla U(x)$ and $\sigma(x) = I$, i.e.,

$$X_{k+1} = X_k - a_k \nabla U(X_k) + \sqrt{2}\, a_k w_k \tag{4.29}$$

where

$$a_k = \frac{c}{\log k} \qquad (k \text{ large}) .$$

Kushner [21] has shown (roughly) that if c is large enough and the sample paths $\{X_k\}$ are bounded with probability one by some device, then X_k converges to S in probability.

Now consider the discretization (4.5) of the Langevin algorithm (4.1) with discretization interval ϵ, i.e.,

$$x_{k+1}^{\epsilon} = x_k^{\epsilon} - \epsilon \nabla U(x_k^{\epsilon}) + \sqrt{2T(k\epsilon)\epsilon}\, w_k . \tag{4.30}$$

Interpolate $\{x_k^{\epsilon}\}$ into $x_{\epsilon}(\cdot)$ with sample paths in $D^r[0,\infty)$ by

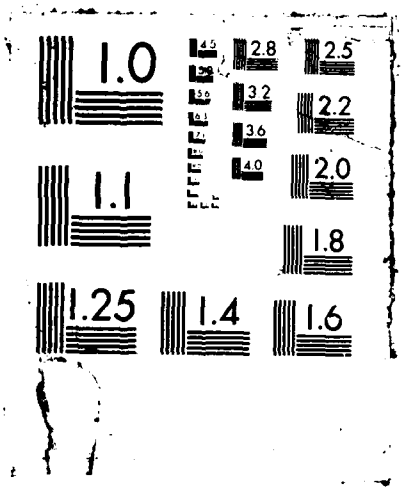
$$x_{\epsilon}(t) = x_k^{\epsilon} \qquad \forall\, (k-1)\epsilon \le t \le k\epsilon , \qquad \forall\, k \in \mathbb{N} .$$

An application of Theorem 4.1 under assumptions (A1), (A2) of Theorem 4.2 shows that

$$x_{\epsilon}(\cdot) \longrightarrow x(\cdot) \quad \text{weakly} \quad (\text{in } D^r[0,\infty)) \quad \text{as} \quad \epsilon \longrightarrow \infty .$$

Suppose in (4.30) we replace the fixed discretization interval ϵ by a_k, and the accumulated time $k\epsilon$ by

AD-A189 382 ANALYSIS OF SIMULATED ANNEALING TYPE ALGORITHMS(U) 2/2
MASSACHUSETTS INST OF TECH CAMBRIDGE LAB FOR
INFORMATION AND D.. S B GELFAND ET AL. MAY 87
UNCLASSIFIED LIDS-TH-1668 AFOSR-TR-87-1916 F/G 12/3 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

$$t_k = \sum_{n=1}^{k-1} a_n ,$$

and define

$$\tilde{X}_{k+1} = \tilde{X}_k - a_k \nabla U(\tilde{X}_k) + \sqrt{2T(t_k)a_k}\, w_k .$$

By L'hopital's rule and the Fundamental Theorem of Calculus

$$T(t_k) = \frac{c}{\log \sum_{n=1}^{k-1} a_n} \sim \frac{c}{\log \int_2^k \frac{1}{\log x}\, dx} \sim \frac{c}{\log k} = a_k$$

as $k \to \infty$. Hence we may write

$$\tilde{X}_{k+1} = \tilde{X}_k - a_k \nabla U(\tilde{X}_k) + \sqrt{2}\, \tilde{a}_k w_k \tag{4.31}$$

where $\tilde{a}_k \sim a_k$. In view of (4.29) and (4.31) it seems clear that we may identify $\{X_k\}$ and $\{\tilde{X}_k\}$ as essentially the same algorithm, and so we can view $\{X_k\}$ as arising from a discretization $\{x_k^\epsilon\}$ of the Langevin algorithm $x(\cdot)$ with a nonstationary discretization interval $\epsilon = a_k$, at least for k large enough. Note that the weak convergence of $x_\epsilon(\cdot)$ to $x(\cdot)$ in $D^r[0,\infty)$ as $\epsilon \to 0$ does not imply that $\{\tilde{X}_k\}$ and $x(\cdot)$ (and presumably $\{X_k\}$ and $x(\cdot)$) have "close" asymptotic measures, from which we might conclude that the convergence of X_k to S in probability as $k \to \infty$ follows from the convergence of $x(t)$ to S in probability as $t \to \infty$ (c.f. [23] for a discussion of asymptotic measures and the relationship to weak convergence). However, the weak convergence of $x_\epsilon(\cdot)$ to $x(\cdot)$ in $D^r[0,\infty)$ as $\epsilon \to 0$ and the convergence of $x(t)$ to S in probability as $t \to \infty$ does provide a certain intuitive basis for the convergence of X_k to S as $k \to \infty$ in probability, which infact Kushner proves.

Using the above interpretation of Kushner's algorithm (4.29) as a certain discretization of the Langevin algorithm (4.1) we now proceed to construct a hybrid annealing/Langevin algorithm. For each $\epsilon > 0$ define an $\mathbb{R}^r$-valued discrete parameter process $\{y_k^\epsilon\}$ as follows. Let

$$y_{k+1}^\epsilon = y_k^\epsilon + \sqrt{2T(k\epsilon)\epsilon}\, m_k^\epsilon\, w_k$$

where $\{m_k^\epsilon\}$ is a sequence of $\{0,1\}$-valued random variables such that m_k^ϵ is conditionally independent of $y_1^\epsilon,\ldots,y_{k-1}^\epsilon$, $w_1,\ldots,w_{k-1}$, and $m_1^\epsilon,\ldots,m_{k-1}^\epsilon$ given y_k^ϵ, w_k, and

$$P\{m_k^\epsilon = 1 | y_k^\epsilon = y\ ,\ w_k = w\} = \exp\left\{ - \frac{[U(y+\sqrt{2T(k\epsilon)\epsilon}\ w) - U(y)]_+}{T(k\epsilon)} \right\},$$

where we use the notation $[a]_+ = \max\{0,a\}$ for $a \in \mathbb{R}$. A calculation shows that $\{y_k^\epsilon\}$ is a Markov chain and

$$P\{y_{k+1}^\epsilon \in A | y_k^\epsilon = y\} = \int_A s_k^\epsilon(y,z)\ dN(y,2T(k\epsilon)\epsilon I)\ (z) + \tilde{\gamma}_k^\epsilon(z)\ \delta(z,A) \quad (4.32)$$

for all $y \in \mathbb{R}^r$ and $A \in \mathcal{B}^r$, where $s_k^\epsilon(\cdot,\cdot)$ is given by (4.8) and $\tilde{\gamma}_k^\epsilon(\cdot)$ is given by the r.h.s. of (4.9) with ϵI replaced by $2T(k\epsilon)\epsilon I$. Comparing (4.32) and (4.7) we see that $\{y_k^\epsilon\}$ like $\{z_k^\epsilon\}$ is an annealing chain driven by white Gaussian noise, except that the noise driving $\{y_k^\epsilon\}$ is *nonstationary* with covariance $2T(k\epsilon)\epsilon I$. Interpolate $\{y_k^\epsilon\}$ into $y_\epsilon(\cdot)$ with sample paths in $D^r[0,\infty)$ by

$$y_\epsilon(t) = y_k^\epsilon \qquad \forall (k-1)\epsilon \leq t < k\epsilon\ , \quad \forall\ k \in \mathbb{N}.$$

In Theorem 4.2 we gave conditions such that

$$z_\epsilon(\cdot) \rightarrow z(\cdot) \quad \text{weakly} \quad (\text{in } D^r[0,\bar{T}]) \quad \text{as} \quad \epsilon \rightarrow 0\ ;$$

minor changes in the proof of Theorem 4.2 show that

$$y_\epsilon(\cdot) \rightarrow x(\cdot) \quad \text{weakly} \quad (\text{in } D^r[0,\infty)) \quad \text{as} \quad \epsilon \rightarrow 0$$

under the same conditions.

Now define an $\mathbb{R}^r$-valued discrete parameter random process $\{Y_k\}$ as follows. Let

$$Y_{k+1} = Y_k + \sqrt{2}\ a_k M_k w_k$$

where $\{M_k\}$ is a sequence of $\{0,1\}$-valued random variables such that M_k is conditionally independent of $Y_1,\ldots,Y_{k-1}$, $w_1,\ldots,w_{k-1}$, and $M_1,\ldots,M_{k-1}$ given Y_k, w_k, and

$$P\{M_k = 1 | Y_k = y\ ,\ w_k = w\} = \exp\left\{ - \frac{[U(y + \sqrt{2}\ a_k w) - U(y)]_+}{a_k} \right\}.$$

By similar reasoning as with $\{X_k\}$ we may view $\{Y_k\}$ as arising from a discretization $\{y_k^\epsilon\}$ of the Langevin algorithm $x(\cdot)$ with a nonstationary discretization interval $\epsilon = a_k$, at least for k large enough. We shall call the algorithm which simulates the sample paths of $\{Y_k\}$ the *hybrid annealing/Langevin algorithm.*

We shall now make a few comments concerning the convergence of the hybrid annealing/Langevin algorithm. The weak convergence of $y_\epsilon(\cdot)$ to $x(\cdot)$ in $D^r[0,\infty)$ as $\epsilon \rightarrow 0$ and the convergence of $x(t)$ to S in probability as $t \rightarrow \infty$

does provide an intuitive basis for the convergence of Y_k to S in probability as $k \to \infty$. This intuition is further bolstered by the convergence of X_k to S in probability as $k \to \infty$. Unfortunately, we have not been able to establish the convergence of $\{Y_k\}$. One approach which might be fruitful is to try to adapt Kushner's proof of the convergence of $\{X_k\}$ (we have not tried this). Our idea which we did not succeed in developing was to try to obtain the asymptotic behavior of the $\{Y_k\}$ process *directly* from the asymptotics of the related $x(\cdot)$ process. This is similar in some respects to the *associated ODE* method used to analyze stochastic approximation algorithms (c.f. [24]), whereby the asymptotics of the stochastic approximation algorithm are obtained from the asymptotics of the "limit" process which satisfies an ordinary differential equation. However in our problem the "limit" process $x(\cdot)$ satisfies the stochastic differential equation (4.1). Without going into details it now appears to us that the nonstationarity of $x(\cdot)$ makes it very difficult (if not impossible) to extend the associated ODE method to prove convergence of $\{Y_k\}$.

It is interesting to compare the 1-step transition probabilities for $\{X_k\}$ and $\{Y_k\}$. Let $N(m,\Lambda)(\cdot)$ be an r-dimensional Gaussian density with mean m and positive definite covariance Λ, i.e.

$$N(m,\Lambda)(\xi) = \frac{1}{(2\pi)^{r/2}(\det \Lambda)^{1/2}} \cdot \exp\left[-(\xi - m, \Lambda^{-1}(\xi - m))/2\right]$$

for all $\xi \in \mathbb{R}^r$. Then we may write

$$P\{X_{k+1} \in A \mid X_k = \eta\} = \int_A f(\eta,\xi)\, d\xi$$

$$P\{Y_{k+1} \in A \mid Y_k = \eta\} = \int_A g(\eta,\xi) d\xi + \tilde{\gamma}(\eta)\, \delta(\eta,A)$$

for all $\eta \in \mathbb{R}^r$ and $A \in \mathcal{B}^r$, where

$$f(\eta,\xi) = N(\eta + a_k \nabla U(\eta), 2a_k^2 I)\ (\xi)$$

$$g(\eta,\xi) = \exp\left\{-\frac{[U(\xi) - U(\eta)]_+}{a_k}\right\} N(\eta, 2a_k^2 I)\ (\xi)$$

$$\tilde{\gamma}(\eta) = 1 - \int g(\eta,\xi)\, d\xi$$

for all η, $\xi \in \mathbb{R}^r$. In Figure 4.1(a) we show a bimodal $U(\cdot)$ defined on $\mathbb{R}$. The points ξ_1, ξ_2, ξ_3 are solutions of $U(\xi) = U(\eta)$ for a fixed η. In Figure 4.1(b) we sketch $f(\eta,\cdot)$. In Figure 4.1(c) we sketch $g(\eta,\cdot)$; we also show the unweighted Gaussian density $N(\eta, 2a_k^2)\ (\cdot)$ and the atom at η with mass $\tilde{\gamma}(\eta)$. These

figures make clear the "local" behavior of the Langevin algorithm versus the "semilocal" behavior of the hybrid annealing/Langevin algorithm as discussed in 4.1.

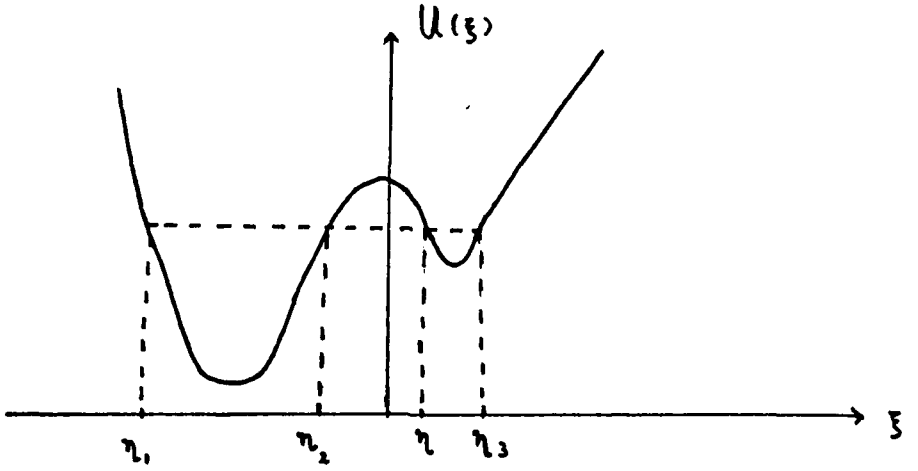


Fig. 4.1(a) Bimodal U(·)

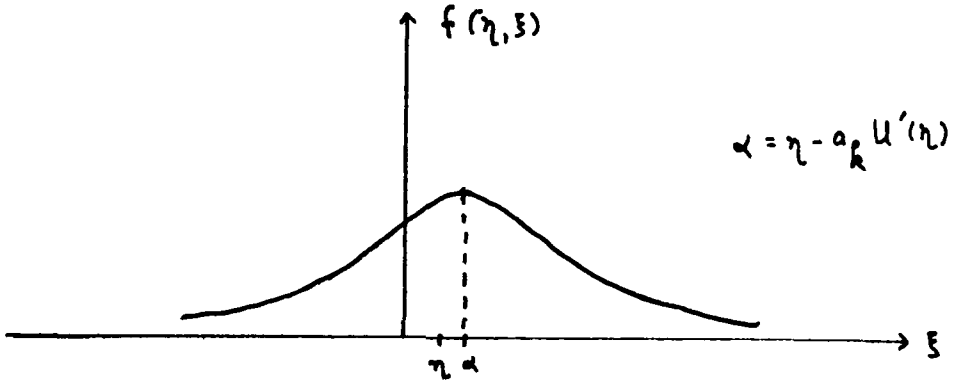


Fig. 4.1(b) Density for X_{k+1} Given X_k

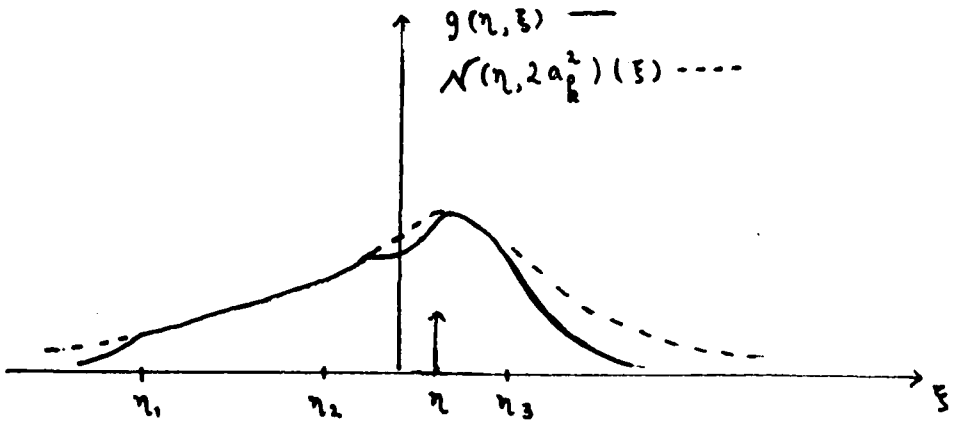


Fig. 4.1(c) Density for Y_{k+1} Given Y_k

CHAPTER V
CONCLUSIONS

5.1 Summary of Results

We summarize the results of this thesis as follows.

(i) We analyzed the rate of convergence in probability of the annealing chain for a special case of an energy function with two local minima. We obtained convergence rates for nonparametric temperature schedules (Theorem 2.8), and also for parametric temperature schedules $T_k = c/\log k$ for $c \geq \Delta^*$ where Δ^* is Hajek's optimal constant (Corollary 2.2). There are two factors which limit the rate of convergence in probability. One factor corresponds to the rate at which the annealing chain makes transitions from one local minimum to the other and back. For temperature schedules $T_k = c/\log k$ this factor dominates whenever $c > \Delta^*$. The other factor corresponds to the rate at which the annealing chain makes its first transition from the strictly local minimum to the global minimum. For temperature schedules $T_k = c/\log k$ this factor is only important when $c = \Delta^*$. We gave explicit expressions for the characteristic time scales associated with each of the rate limiting factors.

(ii) We analyzed the sample path properties of the annealing chain. We gave conditions such that the annealing chain visits the set S of globally minimum energy states with probability one (Theorem 2.9), visits S with probability strictly less than one (Theorem 2.10), and converges to S with probability one (Theorem 2.11).

(iii) We gave a modification of the annealing algorithm so as to allow for noisy measurements of the energy differences which are used in selecting successive states. This is important when the energy differences cannot be measured exactly or when it is simply too costly to do so. We focused on the case when at the k^{th} time step the energy difference between the candidate and current states is measured with additive Gaussian noise with mean 0 and variance σ_k^2. We showed that if $\sigma_k^2 = o(T_k^4)$ then the asymptotic behavior of the modified annealing algorithm is essentially the same as that of the unmodified annealing algorithm (Theorem 2.12, Corollary 2.3).

(iv) We extended the annealing algorithm for optimization on general spaces. We generalized our result on the finite state annealing chain visiting the set S of globally minimum energy states with probability one (Theorem 2.9) to the general state annealing chain visiting a neighborhood of S with probability one (Theorem 3.4), essentially under the conditions that the state space be a compact metric space and the energy function be continuous.

(v) Our most important results concern the relationship between the annealing and Langevin algorithms. We showed that a parametric family of annealing chains driven by white Gaussian noise and interpolated into piecewise constant processes converge weakly to a time-scaled Langevin diffusion (Theorem 4.2). Although both the annealing chain and Langevin diffusion at a fixed temperature have a Gibbs invariant measure, the weak convergence seems to us to be a rather surprising result. Motivated by this convergence result, we proposed a hybrid annealing/Langevin algorithm, whose small time behavior resembles that of the annealing algorithm and whose large time behavior is similar to the Langevin algorithm.

5.2 Open Questions

We list here some questions which naturally follow from our work.

(i) Is their an extension of Theorem 2.8 and Corollary 2.2 on the rate of convergence in probability of an annealing chain with an energy function with two local minima to energy functions with an arbitrary number of local minima? Also, in Theorem 2.8 do the conditions (2.54), (2.55) suggest the kind of regularity in the temperature schedule which guarantees fast convergence (recall that only (2.53) is required for convergence)? Also, in view of the relationship discussed in Chapter 4 between the annealing and Langevin algorithms, is it possible to establish rates of convergence similar to those in Theorem 2.8 and Corollary 2.2 for the Langevin algorithm with a smooth energy function with two local minima?

(ii) Does the general-state annealing chain converge in probability to the set of globally minimum energy states, assuming only that the state space is a compact metric space, the energy function is continuous, and suitable conditions on the temperature schedule?

(iii) Finally and most importantly, does the hybrid annealing/Langevin algorithm converge, and does it indeed improve on the performance of the annealing and Langevin algorithms?

REFERENCES

[1] Ash, R.: *Real Analysis and Probability,* Academic Press (1972).

[2] Billingsley, P.: *Convergence of Probability Measures,* Wiley (1968).

[3] Cerny, V.: "A Thermodynamical Approach to the Travelling Salesman Problem: An Efficient Simulation Algorithm," Preprint, Inst. of Phys. and Biophys., Comenius Univ., Bratislava (1982).

[4] Chiang, T.S., C.R. Hwang and S.J. Shew: "Diffusion for Global Optimization in $\mathbb{R}^n$," Preprint, Institute of Mathematics, Academia Sinica, Tapei, Taiwan (1985).

[5] Dixon, L.C.W. and G.P. Szegö: *Towards Global Optimization,* North Holland (1978).

[6] Doob, J.: *Stochastic Processes,* Wiley (1953).

[7] Feller, W.: *An Introduction to Probability Theory and Its Applications,* (Vol. 1), Wiley (1957).

[8] Geman, S. and D. Geman: "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," IEEE Trans. PAMI 6 (1984), 721-741.

[9] Geman, S. and C.R. Hwang: "Diffusions for Global Optimization," SIAM J. Cntrl. Opt. 24 (1986) 1031-1043.

[10] Gidas, B.: "Nonstationary Markov Chains and Convergence of the Annealing Algorithm," J. Stat. Phys. 39 (1985) 73-131.

[11] Gidas, B.: "Global Optimization via the Langevin Equation," Proc. IEEE Conf. Dec. and Cntrl. (1985).

[12] Golden B. and C. Skiscim: "Using Simulated Annealing to Solve Routing and Location Problems," Naval Res. Log. Quarterly 33 (1986) 261-279.

[13] Grenander, U.: *Tutorial in Pattern Theory,* Brown University (1983).

[14] Hajek, B.: "Cooling Schedules for Optimal Annealing," Preprint, Dept. of Elec. Eng. and Coord. Science Lab., U. Illinois at Champaign-Urbana (1985).

[15] Hajek, B.: "Tutorial Survey of Theory and Applications of Simulated Annealing," Proc. IEEE Conf. Dec. and Cntrl. (1985).

[16] Hammersley, J. and D. Handscomb: *Monte Carlo Methods,* Chapman and Hall (1964).

[17] Hwang, C.R.: "Laplace's Method Revisited: Weak Convergence of Probability Measures," Ann. Prob. 8 (1980) 1177-1182.

[18] Johnson, D.S., C.R. Aragon, L.A. McGeoch, and C. Schevon: "Optimization by Simulated Annealing: An Experimental Evaluation," Preprint (1985).

[19] Kirkpatrick, S., C.D. Gelatt, and M. Vecchi: "Optimization by Simulated Annealing," Science 220 (1983) 621-680.

[20] Knopp, K.: *Theory and Application of Infinite Series,* Hafner (1971).

[21] Kushner, H.: "Asymptotic Behavior for Stochastic Approximations and Diffusion with Slowly Decreasing Noise Effects: Global Minimization via Monte Carlo," Preprint, Div. Applied Math., Brown University (1985).

[22] Kushner, H.: "On the Weak Convergence of Interpolated Markov Chains to a Diffusion," Ann. Prob. 2 (1974) 40-50.

[23] Kushner, H.: *Approximation and Weak Convergence Methods for Random Processes,* MIT Press (1984).

[24] Kushner, H. and D. Clark: *Stochastic Approximation for Constrained and Unconstrained Systems,* Springer Verlag (1978).

[25] Metropolis, M., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller: "Equations of State Calculations by Fast Computing Machines," J. Chem. Phys. 21 (1953) 1087-1091.

[26] Mitra, D., F. Romeo, and A. Sangiovanni-Vancentelli: "Convergence and Finite-time Behavior of Simulated Annealing," Proc. IEEE Conf. Dec. and Cntrl. (1985).

[27] Orey, S.: *Limit Theorem for Markov Chain Transition Probabilities,"* Van Nostrand (1971).

[28] Royden, H.: *Real Analysis,* Macmillan (1964).

[29] Rubenstein, R.: *Simulation and the Monte-Carlo Method,* Wiley (1981).

[30] Tsitsiklis, J.: "Markov Chains with Rare Transitions and Simulated Annealing," Preprint, Lab. for Info. and Decision Systems (1985).

[31] Varadhan, S.R.S.: *Lectures on Diffusion Problems and Partial Differential Equations,* Tata Institute, Bombay (1980).

[32] Aluffi-Pentini, F., V. Parisi, and F. Zerilli: "Global Optimization and Stochastic Differential Equations," J. Opt. Th. Applic. (1985) (to appear).

END

DATE

FILMED

MARCH

1988

DTIC